ClearView: Data Cleaning for Online Review Mining

AMANDA MINNICH, NOOR ABU-EL-RUB, MAYA GOKHALE, RONALD MINNICH, AND ABDULLAH MUEEN

UNIVERSITY OF NEW MEXICO, USA
Mining social media data is important, but data cleaning dominates.

What data scientists spend the most time doing:

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

CrowdFlower 2016 Data Science Report
Mining social media data is important, but data cleaning dominates

Why That's a Problem
Simply put, data wrangling isn't fun. It takes forever. In fact, a few years back, the New York Times estimated that up to 80% of a data scientist's time is spent doing this sort of work.

Here, it's necessary to point out that data cleaning is incredibly important. You can't do the sort of work data scientists truly enjoy doing with messy data. It needs to be cleaned, labeled, and enriched before you can trust the output.

The problem here is two-fold. One: data scientists simply don't like doing this kind of work, and, as mentioned, this kind of work takes up most of their time. We asked our respondents what was the least enjoyable part of their job.

They had this to say:

Note how those last two charts mirror each other. The things data scientists do most are the things they enjoy least. Last year, we found that respondents far prefer doing the more creative, interesting parts of their job, things like predictive analysis and mining data for patterns. That's where the real value comes. But again, you simply can't do that work unless the data is properly labeled. And nobody likes labeling data.

What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

CrowdFlower 2016 Data Science Report
Why focus on review data?

- Full of noise, good testing ground
- Used by people all over the world: huge data set
- Can be mined using NLP, network analysis, and time series algorithms
The review space is full of noise

Josh Collins  October 28, 2015

Cool App! Also try "WILD WALLET" - Money Online! Download "Wild Wallet" Right Now! Do not forget to enter My Bonus Code: 1050157

Jason A. James  April 2, 2015

Gifts happen win our nice. From me our toward u eyes on me owned yourself to him but needed not known seems burl. A brand new automobile

Philip Jimenez  July 12, 2016

Plz help me. Please join refer sponsor ID 1033068 My dream come true... (invite plz 1033068) Oh I dreamed to earn money by myself and purchase apps from it and now I am only 9 and I've purchased Disney Princesses Story Theatre with my earned money. My deepest thank you to the inventor of this app. :)...ok Plz help me. Please join refer sponsor ID 1033068

Dayvid Acosta  May 21, 2014

Yousss XT gg0yz Hsu idd grid gg9 I'd I'd yhh huh hi ICCICC

Amaury Moore  September 14, 2014

TV DVD So she eh disk disk scorn Usk it'll um ICM IFK if ICM TDK do flat ekks using us is TV ng to us HK u do is runs in IDB ICM do DVT egg RSM ox ran LCC RSC

Sheila Narayanasamy  March 23, 2015

Real drum This game is the best game I have played

Jess Denn  February 15, 2016

Kennel confirm um and then we are the new one is not be able to be a lot and then you can be a lot to the last week of the way to be a lot and I have a bit and I have any questions about the way you can you are you are you are you are you are the way you can be in this email is not be in this email is not be in this email is not be in this u next to be able and I can you are not the new Hampshire the new year I am going well as well as well with a lot of
Three types of noise in reviews

1. Syntactic: character/word-level
2. Semantic: sentence-level
3. Rating: review-level

Our goal: automate pipeline to filter noise for improved data mining
ClearView: Related work

- Interactive data cleaning methods:
  - For specific algorithms (S. Krishnan et al, 2016)
  - For use by analysts (S. Krishnan et al, 2016)

- Iterative data cleaning workflow creation (D. Haas et al, 2015)

Our method is the first attempt to automatically filter noisy reviews
Our publicly available dataset

- **tripadvisor**
  - 3,173,346 reviews
  - 46,095 hotels
  - 1,426,252 users
  - 10,000 reviews annotated with 3 sentiment scores

- **Total**
  - 65.5 million reviews
  - 270k hotels
  - 1.1 million apps
  - 26.5 million users

- **Google play**
  - 50,360,257 reviews
  - 1,146,054 apps
  - 23,637,232 users
  - 10,000 reviews annotated with 3 sentiment scores

These and other datasets are available at: http://cs.unm.edu/~aminnich
Contributions: ClearView

- **Create automated pipeline to filter out noise**
- **Identify 3 common types of noise and design cleaning methodologies**
- **Establish ground truth using Amazon Mechanical Turk**
- **Illustrate dramatic improvement in quality of dataset through experimental evaluation**
Roadmap

- Syntactic noise + cleaning methods
- Semantic noise + cleaning methods
- Rating noise + cleaning methods
- Validation and experimental evaluation
Roadmap

- Syntactic noise + cleaning methods
- Semantic noise + cleaning methods
- Rating noise + cleaning methods
- Validation and experimental evaluation
Syntactic noise: What is it?

- Non-ASCII characters
- Non-English reviews
- Spam-related keywords
- Meaningless words
Syntactic noise: Identifying and Cleaning

- Python’s *Printable* function: identified percentage of non-ASCII characters
- Python’s *Enchant* library: identified percentage of words that are in its corpus
- Blacklist: identified percentage of spam-related keywords
- Tunable thresholds allow for user customization
Syntactic noise: special case

Josh Collins  October 28, 2015

Cool App! Also try "WILD WALLET" - Money Online! Download "Wild Wllt" Right Now! Do not forget to enter My Bonus Code: 1050157

Cool ! Also try "WILD WLL" - Mon Online! Downld "Wild Wllt" Right Now! Do not forget to ntr nus d: 1050157.

HIDDEN PHRASES: WALLET, Money, Download, Wallet, enter My Bonus Code

User has left 11 other reviews with this signature
Roadmap

- Syntactic noise + cleaning methods
- Semantic noise + cleaning methods
- Rating noise + cleaning methods
- Validation and experimental evaluation
Semantic noise: What is it?

- Badly formed sentences that are unintelligible
- Valid words used, but in nonsensical way
Semantic noise: Identifying and Cleaning

- Stanford CoreNLP Parser: parses sentences into trees
- Confidence score of parser: implicitly measures semantic correctness of a sentence
- Semantic score of review = confidence scores of sentences normalized by length and averaged
Distribution of semantic scores

I'll give 5 stars if the problem is solved.

B. Booty booty booty booty booty booty
    booty booty booty booty booty booty
    booty booty booty booty booty booty
    booty booty booty booty booty booty
    booty booty booty booty booty booty...
Semantic noise: Scores for unintelligible reviews

The lower the score, the more unintelligible the review

-220.63

Jason A. James  April 2, 2015
5 stars
Gift's happens win our nice. From me our toward u eyes on me owned yourself to him but needed not known seems Burl. A brand new automobile

Amaury Moore  September 14, 2014
5 stars
TV DVD So she eh disk disk scorn Usk it'll urn ICM IFK if ICM TDK do flat eeks using us is TV ng to us HK u do is runs in IDB ICM do DVT egg RSM ox ran LCC RSC

Jess Denn  February 15, 2016
5 stars
Kernel confirm cm and then we are the new one is not able to be a lot and then you can be a lot to the last week of the way to be a lot and I have a bit and I have any questions about the way you can you are you are you are you are the way you can in this email is not be in this email is not be in this email is not be in this u next to be able and I can you are not the new Hampshire the new year I am doing well as well as well with a lot of

-778.99

-478.54
Roadmap

- Syntactic noise + cleaning methods
- Semantic noise + cleaning methods
- Rating noise + cleaning methods
- Validation and experimental evaluation
Rating noise: What is it?

Rating does not match sentiment of text

Sheila Narayanasamy  March 23, 2015
Real drum  This game is the best game I have played

SATHIYA SILAN  March 20, 2015
Good app  Easy for use...nice!

Mohamed Alhamdhi  January 23, 2014
Nice  Nice game easy to play

KRISH AGRAWAL  March 22, 2015
Bug in update  Current update there is calibration problem in time or pace same route same speed is giving different pace

jainuel peterson  March 24, 2015
What the hell has happened to transfers now not available.

Sujit Maan  March 30, 2015
Login problem  My farm in the cloud but I can login my farm PIs help me out
Rating noise: Identifying and cleaning

This process was repeated independently using 2 different sentiment classifiers
Rating noise: Cleaning – Classifier One

- Stanford’s Sentiment Classifier from CoreNLP suite
- State-of-the-art recursive neural tensor net classifier
- Parses sentences into trees, rather than bag-of-words approach
- Accurate, but very slow
Rating noise: Cleaning – Classifier Two

- Naïve Bayes classifier based on [1]
- Handles negation and double negation
- Adds a negated form of the word to the opposing class’s word bank
- Uses unigrams, bigrams, and trigrams

Classifier modification to add new features

- 5-class classification
- Iterative training
- Input and output pipelines with visualizations
- Modified classifiers are publicly available: [https://github.com/amandajean119/sentiment](https://github.com/amandajean119/sentiment)
Roadmap

- Syntactic noise + cleaning methods
- Semantic noise + cleaning methods
- Rating noise + cleaning methods
- Validation and experimental evaluation
Validation: generating ground truth

- Used Amazon Mechanical Turk to generate ground truth
- The sentiment of 10,000 reviews from each dataset were scored by 3 different users
- Ground truth = averaged user score
Evaluation experiments: results

- Naïve Bayes outperforms neural net
- Iterative training improves performance
Evaluation experiments: Design

- **Entire review dataset**
- **Syntactic Filtering**
- **Subset of 60,000 reviews**
  - **Rating noise filtering**
  - **Semantic filtering**
- **Final set of reviews**
- **Subset of 10,000 reviews**
  - **Labeled by Turkers**
  - **Ground Truth**
### Evaluation experiments: results

<table>
<thead>
<tr>
<th></th>
<th># of Reviews</th>
<th>% Filtered after ClearView Pipeline</th>
<th>Correctness before and after filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>TripAdvisor</td>
<td>3,167,036</td>
<td>30%</td>
<td>17.7% =&gt; 59.9%</td>
</tr>
<tr>
<td>Google Play</td>
<td>21,112,036</td>
<td>70%</td>
<td>18.7% =&gt; 36.9%</td>
</tr>
</tbody>
</table>

Given so much noise and a lack of ground truth for most stages of the pipeline, this is a dramatic improvement.

TripAdvisor shows greater improvement due to the extreme noisiness of Google Play’s data.
ClearView pipeline standardized distributions of certain features

These features have been shown to be effective for characterizing anomalous behavior (Minnich et al, WWW2015)
ClearView: Summary

- Create automated pipeline to filter out noise
- Identify 3 common types of noise and design cleaning methodologies
- Establish ground truth using Amazon Mechanical Turk
- Illustrate dramatic improvement in quality of dataset through experimental evaluation
Clearview: Future Work

- Find way to estimate recall after each step
- Evaluate performance of data mining algorithms on cleaned reviews
- Improve performance on Google Play
- Apply to other noisy online datasets
Thank You!

INFORMATION ABOUT THE PROJECT, DATA, AND CODE:
CS.UNM.EDU/~AMINNICH/CLEARVIEW
EMAIL: AMINNICH@CS.UNM.EDU