Strength in Numbers

Can Big Data Eliminate the need for Complex NLP?

Apostolos Filippas
NYU Stern

Theodoros Lappas
Stevens Institute of Technology
Motivation

The ability to know if a sentence is positive or negative is a necessary building block for many applications:

• Review mining & summarization

• Sentiment Analysis (marketing, politics)

• Tweet annotation, recommendation & summarization

…
Solutions

➢ Dozens of papers, mostly on reviews and tweets.

Three main types (and their combinations):

➢ Supervised approaches (requires annotated data)

➢ Lexicon-based approaches (lists of good/bad terms)

➢ **NLP** → by far the best, but complex and expensive
Some Examples

“The food was great”

“Best experience of my life”

“The most boring game ever”

“I don’t recommend the steak”

“The rooms were clean and comfortable”

“The food was horrible and the service was very slow”
Some Examples

“The food was great”

“Best experience of my life”

“The most boring game ever”

“I don’t recommend the steak”

“The rooms were clean and comfortable”

“These are easy….”

“The food was horrible and the service was very slow”
<table>
<thead>
<tr>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Staying longer next time”</td>
<td>“The salad was extra.”</td>
</tr>
<tr>
<td>“the location cannot be beat”</td>
<td>“no apologies were offered”</td>
</tr>
<tr>
<td>“million buck view”</td>
<td>“parking is not very convenient”</td>
</tr>
<tr>
<td>“has a very eclectic feel”</td>
<td>“Did not live up to standards”</td>
</tr>
<tr>
<td>“hats off to the chef”</td>
<td>“Cockroaches everywhere”</td>
</tr>
<tr>
<td>“they go above and beyond”</td>
<td>“the food was meh”</td>
</tr>
<tr>
<td>How about a real challenge?</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>“Minutes to downtown”</td>
<td></td>
</tr>
<tr>
<td>“they have shuttles if needed”</td>
<td></td>
</tr>
<tr>
<td>“check this place out”</td>
<td></td>
</tr>
<tr>
<td>“They leave little chocolates on your pillow.”</td>
<td></td>
</tr>
<tr>
<td>“This was our fourth stay here”</td>
<td></td>
</tr>
<tr>
<td>“we wanted for nothing”</td>
<td></td>
</tr>
<tr>
<td>“there was a charge for self parking”</td>
<td></td>
</tr>
<tr>
<td>“We called reception twice, no response”</td>
<td></td>
</tr>
<tr>
<td>“so much for southern hospitality”</td>
<td></td>
</tr>
<tr>
<td>“jet planes also landing nearby”</td>
<td></td>
</tr>
<tr>
<td>“we waited and waited”</td>
<td></td>
</tr>
<tr>
<td>“I was told the food would take 5 min”</td>
<td></td>
</tr>
</tbody>
</table>
What can we do?

Keep making the lexicons longer…

Keep making NLP more expensive and more complex…

OR…
What can we do?

Keep making the lexicons longer…

Keep making NLP more expensive and more complex…

OR…

Do nothing, let Big Data do ALL the work.
How?

1. Get as many reviews as you can.

2. Split them into sentences.

3. Extract patterns from each sentence.

4. Get the frequency of each pattern in positive and negative reviews.
The rooms were very clean!!!!
The rooms were very clean!!!!

[ the, rooms, were, very, clean ]
The rooms were very clean!!!!

[ the, rooms, were, very, clean ]

[ 1 , rooms, 1 , very, clean ]

Mark Stopwords
The rooms were very clean!!!!

[ the, rooms, were, very, clean ]

[ 1 , rooms, 1 , very, clean ]

[ 1 , rooms, 1 , * , clean ]
[ 1 , * , 1 , very, * ]
[ 1 , rooms, 1 , very, * ]
[ 1 , * , 1 , * , clean ]
[ 1 , rooms, 1 , * , * ]
[ 1 , * , 1 , * , * ]

Extracting Patterns
The rooms were very clean!!!!
The rooms were very clean!!!!

[ the, rooms, were, very, clean ]

[ 1 , rooms, 1 , very, clean ]

[ 1 , rooms, 1 , * , clean ]

[ 1 , rooms, 1 , very, * ]

[ 1 , rooms, 1 , * , * ]

Done, time to start counting!
A Naive Assumption

- All the sentences in a positive review are positive
- All the sentences in a negative review are negative

Positive reviews: 4 or 5 star rating
Negative reviews: 1 or 2 star rating
"Our rooms were very clean." → [1, *, 1, very, clean]

[1, *, 1, very, clean] : [0, 1]

[#, neg, #pos]
“Our rooms were very clean.”

[1, *, 1, very, clean] : [0, 1]

“My suite was very clean.”

[1, *, 1, very, clean] : [0, 2]

The same pattern can emerge from different sentences
The same pattern can emerge from different sentences

1. "Our rooms were very clean."
2. "My suite was very clean."
3. "The hotel is very clean."

The same pattern can emerge from different sentences.
“Our rooms were very clean.”

“My suite was very clean.”

“The hotel is very clean.”

Keep counting until the corpus is exhausted
The rooms were very clean!!!!

Let’s look at another pattern
“Our rooms were very clean.”
"Our rooms were very clean."

"My suite was very small."
+ "Our rooms were very clean."  \[1, * , 1, \text{very}, * \]  \[1, * , 1, \text{very}, * \] : [0, 1]

- "My suite was very small."  \[1, * , 1, \text{very}, * \]  \[1, * , 1, \text{very}, * \] : [1, 1]

- "The hotel is very dirty."  \[1, * , 1, \text{very}, * \]  \[1, * , 1, \text{very}, * \] : [2, 1]
“Our rooms were very clean.”

“My suite was very small.”

“The hotel is very dirty.”

A balanced count
<table>
<thead>
<tr>
<th>Description</th>
<th>Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, *, 1, very, clean]</td>
<td>[3, 237]</td>
</tr>
<tr>
<td>[1, *, 1, very, *]</td>
<td>[499, 486]</td>
</tr>
<tr>
<td>[1, *, 1, very, dirty]</td>
<td>[199, 7]</td>
</tr>
</tbody>
</table>
Can we really trust the counts?

- 500 Vs 502: how confident are you that this is positive?
- What if I have way more positive than negative reviews?
Can we really trust the counts?

➢ 500 Vs 502: how confident are you that this is positive?
➢ What if I have way more positive than negative reviews?

Maybe we should look for a threshold…
Maybe we should find a way to normalize…
Can we really trust the counts?

- 500 Vs 502: how confident are you that this is positive?
- What if I have way more positive than negative reviews?

Maybe we should look for a threshold…
Maybe we should find a way to normalize…

Our initial assumption was that all the sentences in a positive review are positive. Let’s make it a little bit smarter.
A more reasonable Assumption

The valence of a sentence is drawn from an unknown probability distribution (a biased coin).

Ho: *The coin is Fair*

Chi-squared Test (if we have a lot of observations - asymptotic)

Binomial Test (if we have just a few observations - exact)
Valence through Statistical Testing

\[
\begin{bmatrix} 1, & * , & 1, & \text{very}, & \text{clean} \end{bmatrix} : \begin{bmatrix} 3, & 237 \end{bmatrix}
\]

Chi-Stat = 228.15 => Ho rejected at .01 => has valence

\[
\begin{bmatrix} 1, & * , & 1, & \text{very}, & * \end{bmatrix} : \begin{bmatrix} 499, & 486 \end{bmatrix}
\]

Chi-Stat = 0.171 => Ho not rejected at .01 => no evidence for valence

\[
\begin{bmatrix} 1, & * , & 1, & \text{very}, & \text{clean} \end{bmatrix} : \begin{bmatrix} 199, & 7 \end{bmatrix}
\]

Chi-Stat = 178.15 => Ho rejected at .01 => has valence

Valence through Statistical Testing
Does this work?
Our Data

TripAdvisor reviews

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities</td>
<td>161</td>
</tr>
<tr>
<td>Hotels</td>
<td>30,280</td>
</tr>
<tr>
<td>Reviews</td>
<td>4,300,136</td>
</tr>
<tr>
<td>Sentences</td>
<td>38,616,569</td>
</tr>
</tbody>
</table>
Ratings Distribution
Sentence Length Distribution

![Histogram of Sentence Length Distribution](image)
Patterns grow very fast, but very few pass the test

6,829,995 out of 3,665,818,410 patterns accepted (0.18 %)
A User Study

Ignore all sentences with known positive/negative words.

Randomly sample 500 sentences from positive reviews + 500 sentences from negative reviews.

Use students to annotate all 1000 sentences as positive, negative or neutral. Each sentence was annotated by the same 10 students, very small number of disagreements (easy for humans).

583 positive, 371 negative, 46 neutral
A User Study - Results

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>72%</td>
<td>9%</td>
<td>19%</td>
</tr>
<tr>
<td>Negative</td>
<td>3%</td>
<td>74%</td>
<td>23%</td>
</tr>
<tr>
<td>Neutral</td>
<td>14%</td>
<td>20%</td>
<td>66%</td>
</tr>
</tbody>
</table>

+ We get all the **hard** and **very hard** sentences that we saw earlier! (these were all selected from the user-study data)
Let’s make it a little bit harder

“Staying longer next time”
“the location cannot be beat”
“million buck view”
“has a very eclectic feel”
“hats off to the chef”
“they go above and beyond”

“The salad was extra.”
“no apologies were offered”
“parking is not very convenient”
“Did not live up to standards”
“Cockroaches everywhere”
“the food was meh”
How about a real challenge?

“Minutes to downtown”
“they have shuttles if needed”
“check this place out”
“They leave little chocolates on your pillow.”
“This was our fourth stay here”
“we wanted for nothing”

“there was a charge for self parking”
“We called reception twice, no response”
“so much for southern hospitality”
“jet planes also landing nearby”
“we waited and waited”
“I was told the food would take 5 min”
How do we know if we have enough data?
We check:

Pattern Coverage

- \( \text{coverage}(p) = \frac{\text{#distinct sentences that } p \text{ covers}}{\text{# distinct sentences}} \)

- \( \text{weighted coverage}(p) = \frac{\text{#sentences that } p \text{ covers}}{\# \text{sentences}} \)

Sentence purity

- \( \text{purity}(s) = \frac{|\#\text{pos}(s) - \#\text{neg}(s)|}{\#\text{pos}(s) + \#\text{neg}(s)} \)
Coverage becomes harder for larger sentences: get more data or get smarter!

Remember: some sentences are neutral
Near-Perfect Purity, across sentence lengths!
Testing on Unseen Data
(80%-20% training-testing split)
Similar Findings, the algorithm does equally well on known/unknown data
Purity

Near-Perfect Purity, across sentence lengths!
Looking Ahead

A large-scale user study
➢ Compare with NLP & lexicon-based baselines
➢ Easy for English, how about other languages? Sure, as long as we can get enough data (maybe Chinese?)

Can we go beyond single sentences?
➢ Sure, as long as we can get enough data

We need more data.
➢ More reviews
➢ Beyond TripAdvisor
➢ Beyond Hotels
➢ Beyond English

Launch a Live Demo, get the crowd involved.
Thank you!