

# Great Again or Stronger Together? Sentiment Analysis About Book Reviews on Amazon

Jiabao Wu<sup>1</sup>, Siyan Li<sup>1</sup>  
lisiyan6315@foxmail.com

March 8, 2017

---

<sup>1</sup>Central University of Finance and Economics

During the presidential election in the year of 2016, Donald Trump, a fruitful writer, businessman and talk show host, swept adverse remarks away and flushed to the top against his rival, Hillary Clinton, as the 45th American president.

In this election, both of the main candidates have published a book to elaborate their ideas for political propaganda. Donald Trump published *Great Again: How to Fix Our Crippled America*(*GreatAgain* for short) in July 12, 2016 while Hillary Clinton published *Stronger Together: A Blueprint for America's Future*(*Blueprint* for short) in Sep 6, 2016. These two books are both wildly read, commented and frequently referred in their speeches. In another word, there are many interesting and valuable messages implied in book reviews.

Given the messages above, we are aimed at data mining towards book reviews on bestsellers by Trump and Hillary. Our analysis focus on 5 questions listed below.

- How do book readers level Trump and Hillary's bestsellers?
- Which words are mainly used in book reviews on their bestsellers?
- What's the main difference in book reviews between Trump and Hillary?
- How did book readers feel when they give comments?
- What differs mostly between positive comments and negative ones?

The following parts are constructed as follows:

- Collecting book reviews
- Data cleaning
- Drawing word clouds
- Extracting words with high distinction between two books
- Sentiment analysis by VADER between two books
- Extracting words with high distinction between positive and negative reviews
- Sentiment analysis by VADER between positive and negative reviews

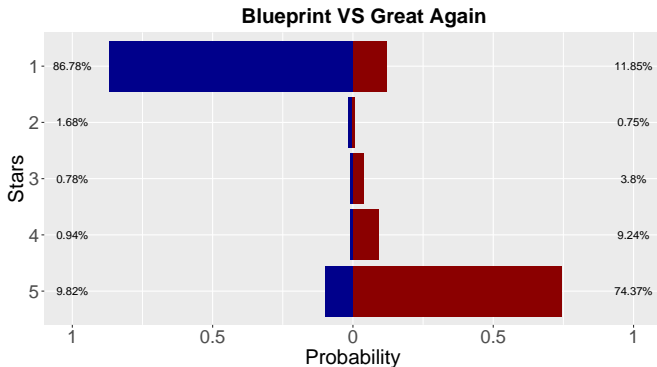
Our research collected book reviews of both candidates' bestsellers on Amazon during US presidential election. *GreatAgain* was published on 12th July, 2016. *Blueprint* was published on 6th September, 2016. We collected all book reviews until the election day (8th November, 2016). An Example of book review is shown below.



Figure: Book review example on Amazon

As is showed in figure above, our data contains 6 sects. Book reviews were cleaned by a filter capable of extracting the number of stars, dates and contexts from Amazon's book reviews. As for other parts, names and formats were excluded for the sake of privacy conservation, and titles were omitted because of duplicate content and the excessive quantity of missing values.

Title	NumberOfReviews	MeanOfRates	HighRates	LowRates
GreatAgain	1342	4.34	1170(87.4%)	169(12.6%)
Blueprint	2556	1.45	295(11.5%)	2261(88.5%)



- Tokenization
- Capitalization conversion
- Noise cleaning(punctuations, numbers and weblinks)
- Stopwords removal
- Remove following four words: 'book', 'read', 'amazon', 'review'
- Stemming and lemmatization

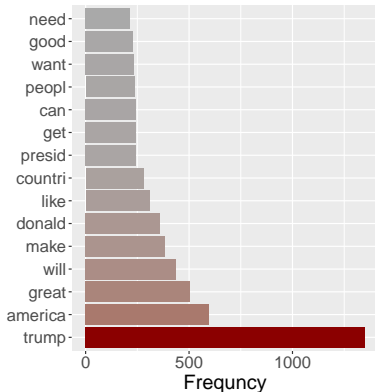
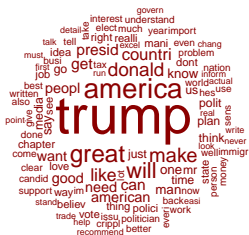


Figure: Book review of *GreatAgain*

The unique words in comments on Trump are 'Great', 'America', 'country', 'preside' and 'good', which are all positive words and closely related to country's future.



## Drawing word clouds

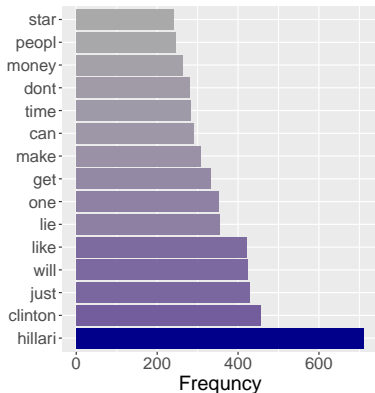
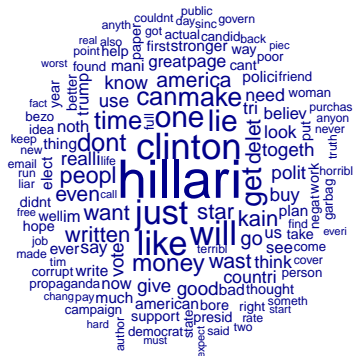


Figure: Book review of *Blueprint*

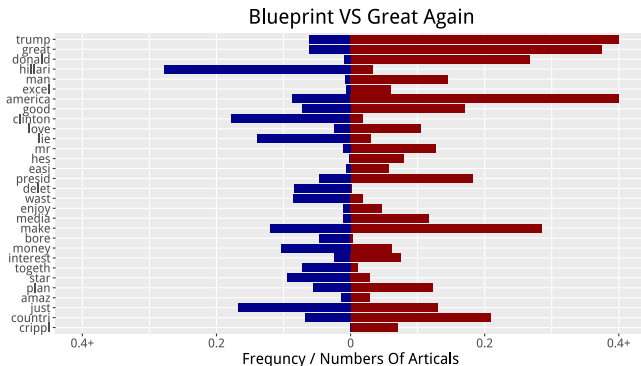
The unique words in comments on Hillary are 'just', 'lie', 'money' and 'dont', which are mostly negative words, attacking Hillary's political affairs.

We consider book reviews as bag-of-words, using word counts as text features and random forest as text classifier in order to classify whether a comment is aimed at Hillary or Trump. The accuracy rate of 10-fold cross validation is 81%, which means our text features are able to extract main differences between different authors.

**Table:** Accuracy & Confusion Matrix

Accuracy:	0.8843		
	Truth	GreatAgain	Blueprint
Predict	GreatAgain	220	25
	Blueprint	30	119

# Extracting words with high distinction between two books



*GreatAgain* features

trump, great, donald, man, excel, american, good, love, mr, hes, easi, presid, enjoy, media, make, interest, plan, amaz, countri, crippl

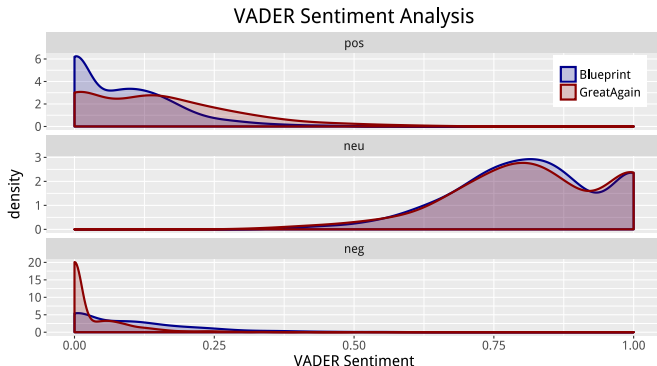
*Blueprint* features

hillari, clinton, lie, delet, money, wast, bore, togeth, star, just

VADER, or Valence Aware Dictionary for sEntiment Reasoning, is a sentiment analysis toolkit, based on Emotional lexicon. This toolkit is proposed in VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. When dealing with microblog-like contexts, VADER outperforms eleven other state-of-practice benchmarks, including LIWC, SentiWordNet, and the General Inquirer. VADER uses a combination of qualitative and quantitative methods to build an exotic emotional vocabulary for Internet slang and its algorithm, based on word-of-bag model, about the same as LIWC. However, VADER is far more expert at discovering word-order sensitive relations between terms.

# Sentiment analysis by VADER between two books

We obtain sentimental evaluation of each review by VADER, each review contains degree of positive, negative and neural emotion. We draw kernel density curve of different emotions between different authors. The result is shown below.

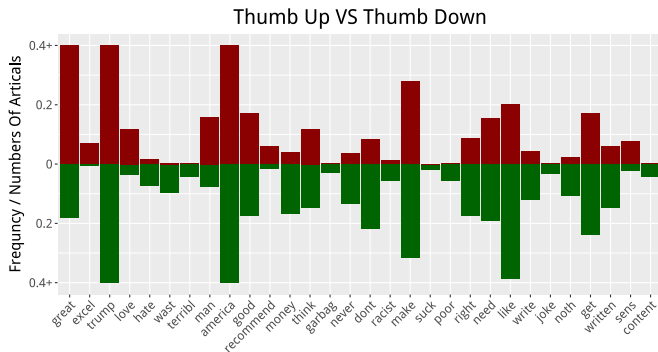


**Figure:** Both candidates' reviews are sentimentally neutralized. However, Trump gets more comments with positive emotion and less with neural and negative ones, while Hillary less positive and more negative.

# Extracting words with high distinction between positive and negative reviews

In this part we aim at exploring difference between positive and negative reviews. We define positive review as those leveled 4 or more stars, and negative review 3 or less stars. We use the bag-of-word model and random forest to extract words with high distinction between positive and negative reviews.

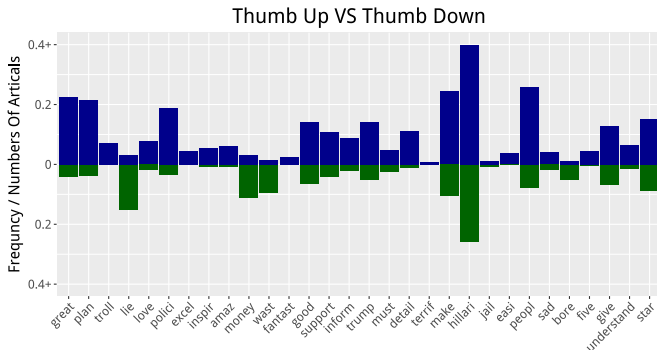
# Extracting words with high distinction between positive and negative reviews



ThumbUp features      great, excel, love, man, good, recommend, sens

ThumbDown features      hate, wast, terribl, money, think, garbag, never, dont, racist, make, suck, poor, right, need, like, write, joke, noth, get, written, content

# Extracting words with high distinction between positive and negative reviews



ThumbUp features

great, plan, troll, love, polici, excel, inspir, amaz, fantast, good, support, inform, trump, must, detail, make, hillari, easi, peopl, sad, five, give, understand, star

ThumbDown features

lie, money, wast, terrif, jail, bore



# Sentiment analysis by VADER between positive and negative reviews

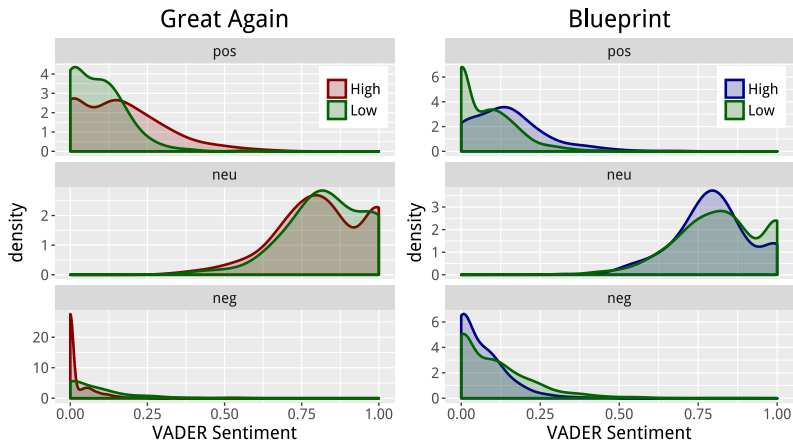


Figure: VADER Sentiment Analysis

## **1.How do book readers level Trump and Hillary's bestsellers?**

On Amazon, most readers gave Trump's bestseller high rates. On the contrary, most readers gave Hillary's bestseller low rates.

## **2.Which words are mainly used in book reviews on their bestsellers?**

The unique words in comments on Trump are Great, America, country, preside and good, most of which are positive words and closely related to country's future. In contrast, the unique words in comments on Hillary are just, lie, money and dont, which are mostly negative words, attacking Hillarys political affairs.

### **3.What's the main difference in book reviews between Trump and Hillary?**

Reviews on the Trump's bestseller used more words that reflect positive emotions and nature of leadership, while reviews on Hillary's bestseller use more words which imply conspiracy theory and reflect negative emotions.

### **4.How did book readers feel when they give comments?**

Compared with the readers of *Blueprint*, the readers of *GreatAgain* used more words to reflect positive emotions, less words to reflect neutral and negative emotions, which shows the readers of *Blueprint* are more exciting than the others.

### **5.What differs mostly between positive comments and negative ones?**