



Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining

Marco Passon*, Marco Lippi^o, **Giuseppe Serra***, Carlo Tasso*

* University of Udine

^o University of Modena and Reggio Emilia

Looking for a Smartphone



Amazon.com: Essential Phone in Halo Gray – 128 GB Unlocked Titanium and Ceramic phone with Edge-to-Edge Display

by Essential 1,743 customer reviews | 712 answered questions

List Price: \$499.99
Price: **\$449.00**
You Save: \$50.99 (10%)

Style: **Only Phone**
Color: **Halo Gray**

With a matte finish to both the titanium and ceramic, this Amazon exclusive Essential Phone has a more industrial look while maintaining its elegant design

Stunning edge-to-edge Quad HD display—the largest screen-to-body ratio of any smartphone. So you get a massive screen, but on a phone that still fits comfortably in one hand

13MP Dual RGB + Mono Rear camera with Portrait Mode and 4K video recording . Get more natural looking shots and incredible high-resolution video

Our Assumption

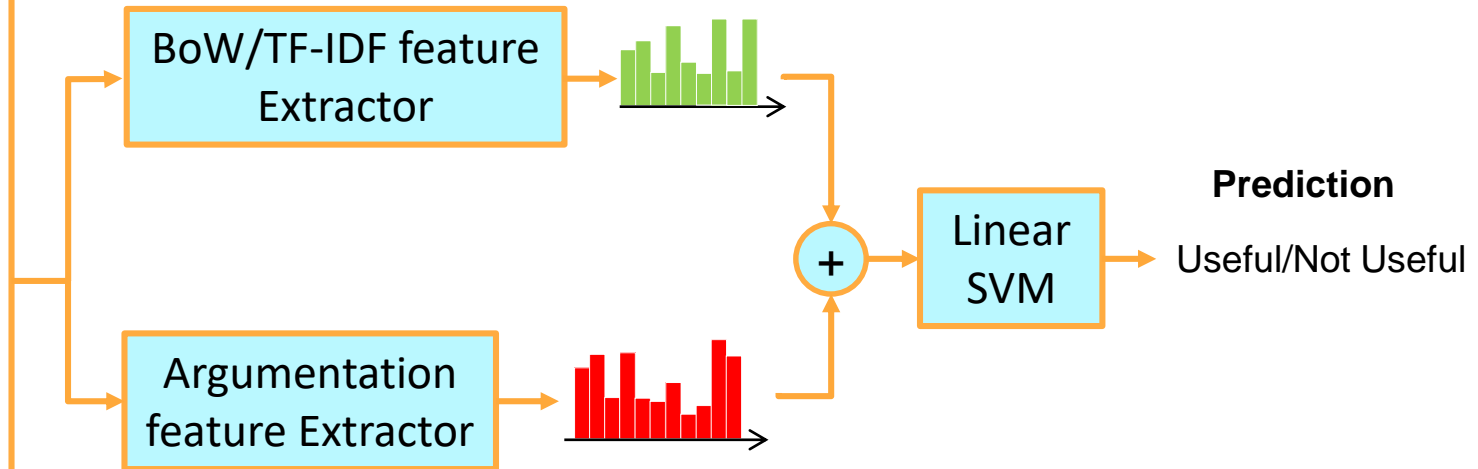
- What **we hope to read** in a review is something that goes beyond plain opinion or sentiment, being rather **a collection or reasons and evidence that support the overall judgment..... In short, we look for argumentative reviews**
- In this work, **we propose a first experimental study** that aims to show how features coming from an **off-the-shelf argumentation mining system can help in prediction** whether a given review is useful.
- **A recent work (Liu et al. 2017*) explores this assumption**, but their study considers a **set of 110 hotel reviews with a manual annotation of arguments**
- **Differently**, in our work **we investigate** the use of **features coming from an automatic system** on a large publicly dataset: **117,000 Amazon Reviews.**

* Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, Hao Wang, "Using Argument-based Features to Predict and Analyse Review Helpfulness", EMNLP 2017

The Proposed Approach

Product Review

Jane Morgan's unpretentious, simple style of singing appealed to me since I was a kid. She put out a lot of records, but is virtually forgotten. It's a shame, because her recordings can serve as the standard for so many modern classics. The only thing I missed on this CD was her recording of "Around the World". Other than that - elegant perfection.



MARGOT System

- **MARGOT is a Websystem that performs argument mining** by exploit a combination of advanced machine learning and natural language processing technique
- Argument Definition (same as Douglas Walton - 2009):
 - **Claim:** a concise statement that directly support or contests a topic
 - **Evidence:** segment text that supports the claim, by bringing a contribution in favour of the thesis that is contained within the claim itself.
- The system was trained on a IBM Research dataset: Debater
 - 547 Wikipedia Articles; 2294 claims and 4690 evidence fact

MARGOT System

Query document

School violence is widely held to have become a serious problem in recent decades in many countries, especially where weapons such as guns or knives are involved. It includes violence between school students as well as physical attacks by students on school staff.



MARGOT

Claim

Evidence

Score_{Claim}

Score_{Evidence}

Score_{Claim}

Score_{Evidence}

MARGOT Pipeline:

- Each document is split in sentences
- Each sentence is processed to produce the Constituency parse tree
- Two classifiers, based on Tree Kernels, detect if a sentence contains claims or evidence facts.

Our Argumentation Features

Product Review

Jane Morgan's unpretentious, simple style of singing appealed to me since I was a kid. She put out a lot of records, but is virtually forgotten. It's a shame, because her recordings can serve as the standard for so many modern classics. The only thing I missed on this CD was her recording of "Around the World". Other than that -- elegant perfection.



MARGOT

Claim	Evidence	Argument (Claim U Evidence)
$Score_{Claim}$	$Score_{Evidence}$	$Score_{Argument}$
$Score_{Claim}$	$Score_{Evidence}$	$Score_{Argument}$
$Score_{Claim}$	$Score_{Evidence}$	$Score_{Argument}$
$Score_{Claim}$	$Score_{Evidence}$	$Score_{Argument}$
$Score_{Claim}$	$Score_{Evidence}$	$Score_{Argument}$



Argumentation features

For each category (Claim, Evidence, Argument) we compute:


- Average (3 features)
- Maximum (3 features)
- N. sentences with score > 0 (3 features)
- Percentage of sentences with score >0 (3 features)

Experimental Evaluation

Amazon Product Dataset

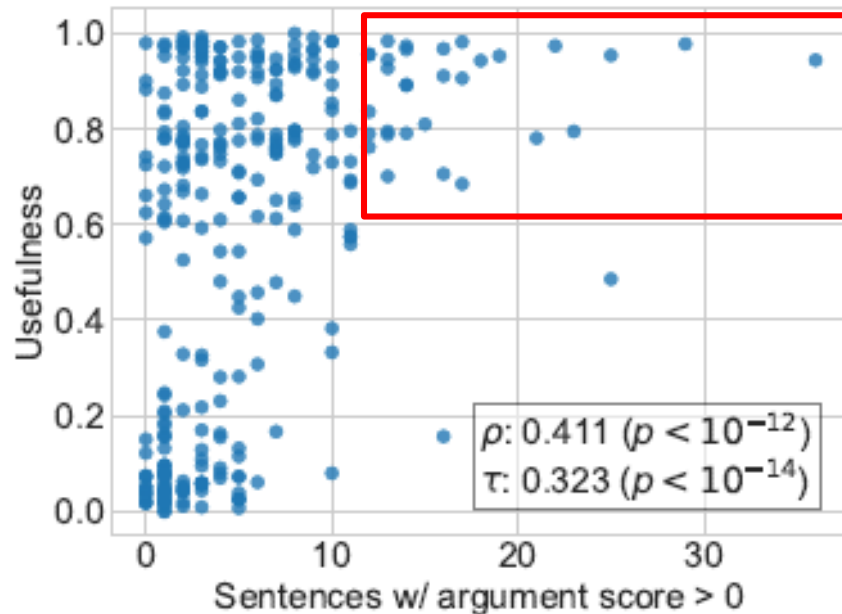
- **Amazon Product Dataset** contains 142.8 million of product reviews spanning May 1996 – July 2014*
- **We select three categories** (CDs and Vinyl, Electronics, TV and Movies) and **we extract**, for each category, **39000 reviews** having at least 75 “helpful” scores.
- **A review is labeled “useful”**, if the ratio between the two numbers is > 0.7

```
{
  'reviewerID': 'A3IYA4H79ISAEH',
  'asin': 'B00BFDHV9E',
  'reviewerName': 'JH',
  'helpful': [76, 88],
  'reviewText': 'I really like the smaller size. I have had
  Sony and Samsung as well as off-brand DVD players in the past
  that were slow to boot up and load CDs. This thing is really
  fast. Loads immediately.',
  'overall': 5.0,
  'summary': 'Works Like It Should',
  'unixReviewTime': 1363219200,
  'reviewTime': '03 14, 2013'
}
```

The Amazon logo, consisting of the word 'amazon' in a sans-serif font with a curved arrow underneath it.

Argumentation vs helpfulness

- Category “CDs and Vinyl” (a random subset of 200 reviews)



- A low number of sentences that contain a claim or an evidence does not necessarily mean that the review is useless
- A review with a high number of sentences containing a claim or an evidence is most likely a useful review

Experimental Results

The experiment has been conducted classifying reviews using:

- **M**: only argumentative features
- **BoW**: only Bag of Words features
- **BoW + M**: combination of Bag of Words and Argumentative features
- **TF-IDF**: only TF-IDF features
- **TF-IDF + M**: combination of TF-IDF and Argumentative features

Metrics: Accuracy (A), Precision (P), Recall (R) and F1 Score (F_1)

Category	Data	A	P	R	F_1
CDs and Vinyl	M	.600	.544	.772	.638
	BoW	.756	.716	.769	.742
	BoW + M	.784	.744	.799	.771
	TF-IDF	.769	.736	.767	.752
	TF-IDF + M	.787	.751	.797	.773
Electronics	M	.583	.529	.744	.618
	BoW	.676	.639	.656	.648
	BoW + M	.689	.640	.714	.675
	TF-IDF	.672	.651	.612	.631
	TF-IDF + M	.689	.649	.684	.666
Movies and TV	M	.564	.517	.792	.625
	BoW	.745	.705	.748	.726
	BoW + M	.773	.741	.767	.754
	TF-IDF	.757	.719	.761	.740
	TF-IDF + M	.777	.739	.784	.761

- **Bag of Words/TF-IDF with argumentative features achieve the best F1 score for each category**

Some Examples #1



- Product Review:

Apple products seemed to be revered as near sacred by Gen Xers. I frankly agree that the beautiful and high-quality surfaces on Apple products is worthy of preservation. This case snaps on easily, fits perfectly, weighs little and does a great job of protecting my Macbook from scratches and mars, even on an airline security conveyor belt.

Prediction

TF-IDF	TF-IDF + M	GT
Not useful		Useful

Some Examples #1



- Product Review:

*Apple products seemed to be revered as near sacred by Gen Xers. I frankly agree that **the beautiful and high-quality surfaces on Apple products is worthy of preservation. This case snaps on easily, fits perfectly, weighs little and does a great job of protecting my Macbook from scratches and mars, even on an airline security conveyor belt.***

Prediction

TF-IDF	TF-IDF + M	GT
Not useful	Useful	Useful

Some Examples #2



- Product Review:

[...] The overrated Neil Gaiman's fantasy nightmares don't even try to make sense; pointless punches are pulled on shallow cartoon characters. The immature Doctor can't shine, stuck with griping harpies. Boo-hoo, Pond leaks. Who cares? Pond's loathsome, "Are we there yet?" of Season Five set the tone for Season Six. [...]

Prediction

TF-IDF	TF-IDF + M	GT
Useful		Not useful

Some Examples #2



- Product Review:

[...] The overrated Neil Gaiman's fantasy nightmares don't even try to make sense; pointless punches are pulled on shallow cartoon characters. The immature Doctor can't shine, stuck with griping harpies. Boo-hoo, Pond leaks. Who cares? Pond's loathsome, “Are we there yet?” of Season Five set the tone for Season Six. [...]

Note: TF-IDF technique has lower performance on long reviewers; this effect is limited by when using argumentation features. Since in this case there are not argumentation sentences, the prediction of our approach is “Not Useful”.

Prediction

TF-IDF	TF-IDF + M	GT
Useful	Not useful	Not useful

Some Examples #3



- Product Review:

I love this product! The price is amazing. It takes a little bit long to boot and the touch screen is a little awkward but overall AMAZING. BUY IT!!

Prediction

TF-IDF	TF-IDF + M	GT
Not Useful		Not useful

Some Examples #3



- Product Review:

I love this product! The price is amazing. **It takes a little bit long to boot and the touch screen is a little awkward but overall AMAZING. BUY IT!!**

Prediction

TF-IDF	TF-IDF + M	GT
Not Useful	Useful	Not useful

Note: Even if there is an argumentation sentence the rest is useless.



Thanks