From Words to Wisdom: Using spaCy for Information Extraction in Customer Reviews

CHATER Oumaima[1]

[1] *AI research student,CentraleSupélec,Paris-Saclay University,France.*
Email: oumaima.chater@student-cs.fr

September 18, 2023

In a world filled with online customer reviews, it can be challenging to find the valuable information you need. This article is all about using a special tool called spaCy to help us make sense of those reviews. With spaCy, we can discover important insights hidden in customer feedback. We'll learn how people feel about products and pick out important details. So, join us as we explore how spaCy works its magic and turns ordinary words into useful wisdom when it comes to customer reviews.

1 Introduction

In our digital age, online reviews are the guiding stars of consumer decision-making. They're like the treasure maps to find the perfect products in the vast e-commerce universe. But, hold on a minute – among all the glittering praise and starry ratings, there's a hidden treasure trove of crucial information. It's like searching for buried treasure, only these gems aren't gold doubloons; they're the insights into product shortcomings.

Picture this: you're eyeing that shiny new tablet. It promises to be your trusty sidekick for work and play. But is it all sunshine and rainbows, or does it have some dark corners you need to know about? That's where we come in.

In this article, we're embarking on an adventure into the world of natural language processing with our trusty companion,spacy, [1]. Our mission is clear: we're diving deep into the ocean of online reviews, particularly those focusing on tablets. Tablets are like modern-day magic wands, but like any wizard's tool, they're not without their quirks.

Our quest is twofold. First, we need to decipher the wild and wonderful language that folks use to describe tablet problems. Imagine translating a secret code, but instead of ancient hieroglyphs, it's emojis and internet slang. Tricky, right? Second, we need to sort and categorize these problems, like organizing a chaotic treasure chest, to help both tablet-makers and potential buyers.

To tackle this challenge, we've enlisted the help of spaCy, a versatile language wizard. By the end of this journey, you'll not only understand what spaCy can do but also how to use its magic to uncover hidden tablet troubles in customer reviews.

Get ready for an exciting expedition into the realm of online reviews. SpaCy is our trusty guide, leading us to the undiscovered tales of tablet hiccups. We're about to shine a light on areas for improvement, offering a roadmap to enhance the tablet experience. So, grab your adventurer's hat and let's set sail!

2 Background

In today's fast-paced world of online shopping, where a world of products is just a click away, the impact of customer reviews is monumental. These digital stories play a crucial role in helping us decide which products to choose, ensuring they're reliable, high-quality, and a good value for our money.

Now, imagine having a tech-savvy friend who can help you make sense of all those reviews. That's where Natural Language Processing (NLP) steps in. Think of it as a super-smart language expert that combines computer magic with human language. It can understand what people are saying in their reviews, even picking up on their feelings and intentions.As defined by IBM [2],"NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning".

But here's the real gem: NLP helps us dig deep into those reviews, uncovering precious information that's often hidden from regular analysis. It's like having a treasure map to find out not only what customers love about a product but also where it might have some issues. For businesses, this treasure trove of insights is gold, guiding them to make their products even better.

And for us, the shoppers? NLP is our trusty guide through the maze of online reviews. It shows us the way by decoding what people really think, highlighting what's great about a product, and gently pointing out its flaws. It's like having a GPS for smart shopping in the digital marketplace.

3 The Challenge of Customer Reviews

Customer reviews, as described in [3],present a multifaceted challenge for information extraction. The diversity in how customers articulate their experiences is striking. Some reviews are succinct, offering only a glimpse of their thoughts, while others sprawl into extensive narratives, overwhelming readers with details. Within this spectrum, there are reviews that contain misspellings, grammatical errors, or ambiguous language, making it challenging to discern their true meaning.

Furthermore, things get even trickier because some reviews don't just talk about product issues; they also mix in personal preferences or outside factors. It's like trying to find a needle in a haystack while the haystack itself is made of different materials. Figuring out which feedback is helpful and which is just personal taste can be a real challenge. This complexity creates a situation where important information about product failures can get lost in the shuffle of reviews.

Traditional methods for dealing with this diversity struggle to keep up. That's why we're suggesting the use of machine learning models. These are like super-smart detectives that can spot patterns, pick out the important stuff, and tell the difference between different opinions. By teaming up with these models, we're on a mission to make it easier to find the failure-related information in the vast world of customer reviews. It's like having a trusty guide to navigate the jungle of opinions and get to the heart of what really matters.

4 Information extraction with spaCy

In this section, we delve into the remarkable capabilities of spaCy in the realm of Natural Language Processing (NLP),and we get into some more technical details. SpaCy stands as a robust toolkit equipped to unravel the mysteries of text data.

4.1 Overview of spaCy:

SpaCy, a leading NLP library, empowers us to harness the power of language. It provides a wide range of tools and functionalities that simplify the complex task of processing and analyzing text. With its efficiency and versatility, spaCy has become a cornerstone in the world of NLP.

4.2 Entity Recognition:

One of spaCy's prominent features is entity recognition. It excels in identifying and categorizing entities within text, be it the names of people, organizations, locations, or specific product mentions. This capability proves invaluable when mining customer reviews for mentions of products or identifying key actors in the review landscape.

4.3 Sentiment Analysis:

SpaCy enables sentiment analysis, allowing us to gauge the emotional tone of customer reviews. By discerning sentiments, such as positive, negative, or neutral, we can gain a deeper understanding of customer perceptions and identify sentiments associated with product failures or successes.

4.4 Topic Modeling:

Another strength of spaCy lies in its ability to perform topic modeling. This technique allows us to extract underlying themes or topics within a corpus of text data. By applying topic modeling to customer reviews, we can unveil recurring subjects or issues, shedding light on common areas of concern or satisfaction. In the following sections, we will explore practical applications of spaCy's capabilities in the context of customer review analysis, demonstrating how it transforms the analysis of text data into a source of actionable insights.

5 Approach and Methodology

In this section, we delve into the intricacies of our unique approach and methodology for leveraging spaCy, a powerful natural language processing (NLP) tool, in the task of information extraction from customer reviews. Our methodology encompasses several crucial steps, each meticulously designed to maximize the accuracy and efficiency of our extraction process.

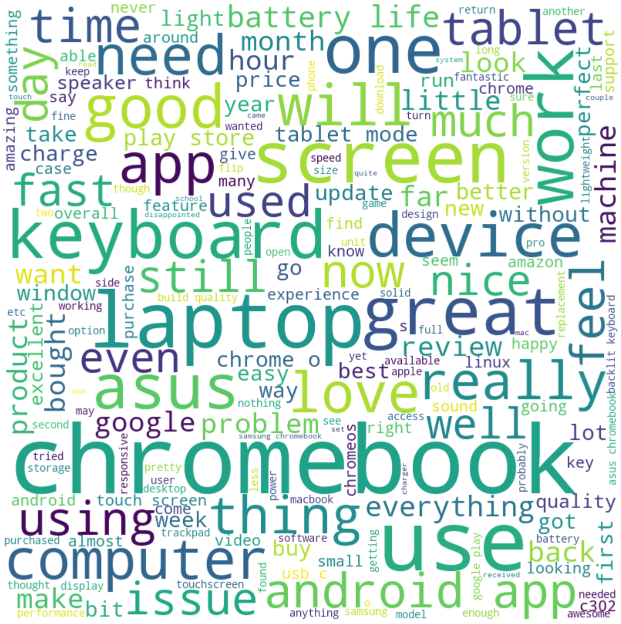
5.1 Data collection and preparation:

We began by collecting a dataset from an e-commerce site, consisting of reviews about a single tablet model. I manually labeled nearly 300 reviews to identify key entities. Before analysis, we cleaned the data by removing noise and ensuring uniform formatting to make it compatible with spaCy.

5.2 Data analysis:

5.2.1 Word Cloud Analysis

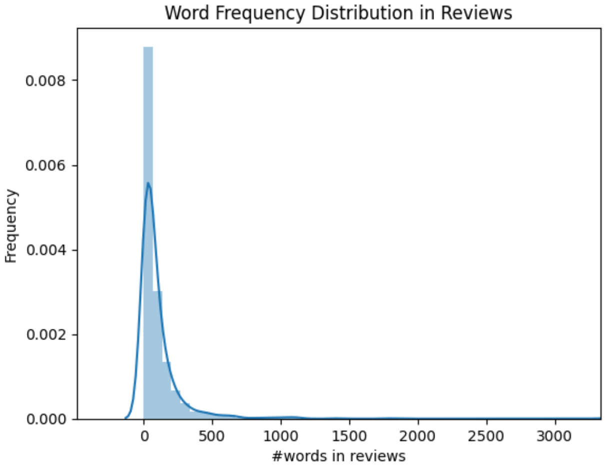
We created a word cloud to visualize the most frequently occurring words within the reviews. This provided a quick glimpse into the common themes and sentiments expressed.



As we can see above, the word cloud highlighted that terms such as "screen," "keyboard," "battery," and "speakers" emerged prominently, indicating that these components were frequently mentioned in customer reviews.

5.2.2 Word Frequency Plot

To delve further, we generated a plot showcasing word frequency across reviews. This graphical representation helped identify patterns and trends in the language used by customers when discussing the tablet.



These visualizations played a crucial role in our exploratory data analysis, serving as invaluable compasses guiding us through the intricate terrain of customer reviews. They allowed us to:

- **Identify Key Themes**: The word cloud immediately drew our attention to the most prevalent topics within the reviews. This insight was pivotal in understanding which aspects of the tablet garnered the most attention and discussion among consumers.

- **Spotlight Components of Interest**: By singling out terms like "screen," "keyboard," "battery," and "speakers," we honed in on the components that were of particular significance to users. This knowledge informed our subsequent information extraction efforts.
- **Unearth Language Patterns**: The word frequency plot illuminated language patterns and recurring terms, shedding light on how customers expressed their experiences and concerns. This comprehension was instrumental in designing effective information extraction strategies.

5.3 Training the NER Model:

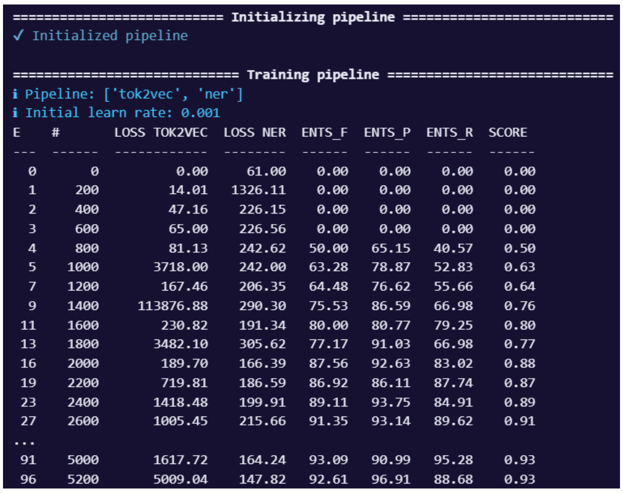
Our journey into information extraction commenced with the training of a custom NER (Named Entity Recognition) model. NER is a subtask of NLP that focuses on identifying and categorizing named entities within text. In our context, named entities could be specific product components (e.g., "screen" or "battery") and failure modes (e.g., "malfunction" or "overheating"). We initiated the training process by executing the following command:

```
!python -m spacy train config.cfg --output
./ --paths.train ./training_data.spacy
--paths.dev ./training_data.spacy
```

During training, the model learned to recognize and categorize entities related to tablets and similar devices, distinguishing between components and failure modes. This custom-tailored model became the cornerstone of our information extraction efforts.

6 Results

In this pivotal phase of our analysis, we unveiled a trove of valuable insights from our information extraction efforts, and we've even captured these insights in a screenshot of the trained model's scores for your reference:



```
========================= Initializing pipeline =========================
✓ Initialized pipeline

========================== Training pipeline ==========================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     61.00    0.00    0.00    0.00    0.00
  1     200         14.01   1326.11    0.00    0.00    0.00    0.00
  2     400         47.16    226.15    0.00    0.00    0.00    0.00
  3     600         65.00    226.56    0.00    0.00    0.00    0.00
  4     800         81.13    242.62   50.00   65.15   40.57    0.50
  5    1000       3718.00    242.00   63.28   78.87   52.83    0.63
  7    1200        167.46    206.35   64.48   76.62   55.66    0.64
  9    1400     113876.88    290.30   75.53   86.59   66.98    0.76
 11    1600        230.82    191.34   80.00   80.77   79.25    0.80
 13    1800       3482.10    305.62   77.17   91.03   66.98    0.77
 16    2000        189.70    166.39   87.56   92.63   83.02    0.88
 19    2200        719.81    186.59   86.92   86.11   87.74    0.87
 23    2400       1418.48    199.91   89.11   93.75   84.91    0.89
 27    2600       1005.45    215.66   91.35   93.14   89.62    0.91
...
 91    5000       1617.72    164.24   93.09   90.99   95.28    0.93
 96    5200       5009.04    147.82   92.61   96.91   88.68    0.93
```

Our NER model achieved remarkable performance in identifying and categorizing entities within customer reviews, as evidenced by the following metrics:

- **F1-Score** ($ENTS_F$): The F1-score reached 0.93, indicating robust entity recognition performance. This high score reflects the model's ability to accurately identify and classify relevant entities within the text.
- **Precision** ($ENTS_P$): Our model demonstrated exceptional precision at 96.91%, signifying a low rate of false positives. This precision is crucial, as it ensures that the identified entities are highly reliable.
- **Recall** ($ENTS_R$): The model achieved a remarkable recall rate of 88.68%, indicating its capacity to capture a significant portion of relevant entities present in the reviews. This high recall ensures that the model doesn't miss important information.
- **Overall Composite Score** ($SCORE$): The combined performance of our model, reflected in the overall composite score of 0.93, underscores its effectiveness in extracting key information from customer reviews.

7 Limitations and Challenges

In our quest to develop cutting-edge technology, we've encountered several hurdles and limitations. Let's uncover the challenges we face as we work to make our project a reality.

7.1 Limitations:

- **Data Quality**: Just like cooking with fresh ingredients makes a better meal, having high-quality data is crucial for our project. If our data is incomplete or unreliable, it can affect the accuracy of our results.
- **Scaling Up**: Imagine trying to serve a banquet with a small kitchen. Scaling up our project to

handle large amounts of data is a bit like that challenge. It's not just about technology; it's also about managing resources efficiently.
- **Language Ambiguity**: Think about how one word can mean different things depending on the context. Our project has to grapple with this language complexity, and it's not always straightforward.

7.2 Challenges:

- **Ethical Roadblocks**: Ensuring the project is ethical and unbiased is a top concern. We need to be vigilant about identifying and correcting potential biases in both our data and the way our algorithms work to ensure fairness.
- **Staying Current**: Language evolves, just like fashion trends. Keeping our technology up-to-date with the latest words, phrases, and linguistic shifts is a constant challenge.
- **Winning Hearts and Minds**: It's one thing to create a fantastic tool, but getting people to embrace it can be as challenging as making the tool itself. We need to convince users that our technology is trustworthy and valuable.

8 Future Directions

Looking ahead, there are several exciting possibilities for further research and enhancements in information extraction:

- Expanding Data Sources: Exploring a wider range of data sources beyond customer reviews, such as social media, forums, or surveys, to gather even more valuable insights.
- Enhanced Models: Investigating advanced machine learning models and techniques that can improve the accuracy and efficiency of information extraction.
- Multilingual Capabilities: Adapting information extraction methods to handle reviews and feedback in multiple languages, catering to a global audience.
- Real-time Analysis: Developing real-time analysis capabilities to provide businesses with immediate feedback and insights from customer interactions.
- Integration with Decision-Making: Integrating extracted information directly into decision-making processes for businesses, aiding in product development and customer support.

These future directions promise to further enhance our ability to extract meaningful information from a diverse range of sources, opening new avenues for understanding and decision-making.

9 Acknowledgments

I would like to express our sincere appreciation to **Zhiguo Zeng** for his invaluable guidance, advice, and supervision throughout the course of this study. His expertise and mentorship have played a pivotal role in shaping the direction and outcomes of this research.

Zhiguo Zeng's unwavering support and insightful contributions have been instrumental in my journey toward understanding and mastering the intricacies of information extraction. I am deeply grateful for his dedication to this study case and his kindness.

References

[1] spacy
`https://spacy.io/`

[2] IBM
`https://www.ibm.com/topics/natural-language-processing`

[3] Zhiguo Zeng,Jie Liu,Jean Meunier-Pion *«Big Data Analytics for Reputational Reliability Assessment Using Customer Review Data»*.01/2021.

[4] Bing Liu *«Sentiment Analysis and Subjectivity*.2010

[5] Nan Hu, Ting Zhang, Baojun Gao, Indranil Bose *«What do hotel customers complain about? Text analysis using structural topic model»*.2019

[6] Peter D.Turney *«Thumbs Up or Thumbs Down?Semantic Orientation Applied to Unsupervised Classification of Reviews»*

[7] Hou, T., B. Yannou, Y. Leroy, and E. Poirson *«Mining changes in user expectation over time from online reviews». Journal of Mechanical Design*.2019