Comparative Product Reviews on Table Tennis Rubbers

Team 15: Winnie Yeung, Jessie Zhang, Pooja Umathe, Eyan Yeung, Charlotte Giang

Table of Contents

1. Introduction

When we shop, we very often find ourselves debating between similar items within the same brand or across different brands. When faced with decisions like these, a natural resource to turn to is searching for comparisons online, e.g. Googling “Amazon echo vs. Google home”, or “Toyota Corolla vs Honda Civic”. This usually brings us into a deep wormhole of forum discussions on comparisons from people who have had experience with both the items in question. This has advantages over comparing individual product reviews separately, as customers who have had experiences with both products have a better first-hand feel of the pros and cons of each. This search process, however, can often be stressful and the shopping experience devolves into a large time sink.

Our goal is to devise an end-to-end comparative tools that comprises natural language processing models that can perform comparative opinion mining on product reviews and forum discussions, and synthesize that information to facilitate decision making for customers in the shopping process.

1.1 Business Applications

The project targets both merchants and customers. For merchants, it is useful to gather customers’ reviews on their products, especially in comparison with the products of their competitors, and what features of their products are outstanding or need improvements. This information informs manufacturing, marketing, customer service, and other key aspects of the business, helping merchants to use their resources more effectively to attract new customers and ensure customer loyalty.

For shoppers, we envision that this project can greatly ease the shopping experience for people who may be debating between different items and eliminate the massive time sink that may go into the decision process. Furthermore, our work would provide a basis for more elaborate product recommendations, thereby also benefiting the companies and online platforms selling these products.

1.2 Scoping

Knowing that product reviews are a broad category, we have narrowed the scope to a specific product - rubber sheets for table tennis paddles - for the purpose of proof of concept. This choice is good for proof of concept since there are generally objective physical features players compare between rubbers, e.g. hardness, spin, speed, weight, etc, as well as subjective properties related to “feel”, e.g. “easier for looping”, or “better for a short game”. This would allow us to explore models for various types of comparison. The methods developed, however, should be fully general and can be applied to any product of interest.

1.3 Literature Review

We found limited literature in comparative opinion mining. While extensive research has been done on text mining, comparative opinion mining has been an untapped branch of natural language processing. The majority of research on comparative opinion approaches uses natural language processing, with a focus on comparative statements in general instead of comparative opinions, meaning the models in use fail to identify reviewers' preferred entities (Varathan & Crestani, 2017). Some scholars have built graphical models to extract comparative relations between two entities, but not a joint model to extract both relations and entities being compared (Xu, Liao, Li, & Song, 2010).

1.4 Methodology

Our project sheds light on comparative opinions of table tennis rubbers, an untapped product in the sports industry, and opens the possibility for future research on various consumer products. More spceifically, the project implements state-of-the-art comparative opinion mining techniques to a specific domain and create an end-to-end product for consumers who are hoping to compare table tennis rubbers in a more effective manner. We have broken down steps of implementation into smaller components: namely identification of aspects of comparison and identification of two entities in comparison. This comparative opinion mining approach uses state-of-the-art natural language processing models and extracts comparative opinions from comparative reviews.

In terms of end-to-end productionalization, we carried out the following steps in sequence:

1. Data collection: This step involves web-scraping online forums and using data wrangling techniques to extract clean text from relevant tags and store the data in a database
2. Data preprocessing: It involves removing non-ASCII values, special characters, HTML tags, emojis, and stop words
3. Classification: Create a classification model for classifying comparative sentences into categories based on objects, attributes, and comparative direction. A preliminary step to create a ground truth dataset is required as well.
4. Identifying aspects of comparison: using a variety of unsupervised and supervised techniques to help narrow down the attributes (hardness, speed, etc) that we want to compare the rubbers for. This includes finding the most popular n-gram, generating comparative word clouds, and k-means clustering.
5. Entity tagging: we identify the subjects of the comments using a lexicon-based approach.
6. Overall Aggregation & Display: We aggregate all the results and store them in an SQLite database prior to passing the result for the frontend dashboard.

2. Data Analysis & Computation

2.1 Data Collection

We have scraped around 16,000 forum discussion threads and posts from 3 table tennis forums such as tabletennisdaily, ooaktabletennis, and MyTabletennis.net. Assuming each of the 16,000 threads has an average of 20 messages, we would have 320,000 messages as a data set to perform comparative opinion analysis. On a sentence level, it approach 1 million sentences, which provide us a plethora of phrases for analysis.

This information is stored in MongoDB database with each document containing information on:

- '_id': mongoDB unique ID
- 'meta_idx': the forum index for the thread
- 'title' : thread title
- 'url': the url ending
- 'author': the author username
- 'num_replies': how many posts the thread has
- 'replies': posts under the thread, each post has information on:
  - message_id: id of the post
  - user_name: user name of the person posting the post
  - reply_msg_id: if the poster is replying to a post
  - post_date: MM-DD-YYYY format of the date of the post
  - clean_text: the text of the post

One advantage of this dataset is that it provides extensive text information on discussions related to table tennis, including our subject of interest: comparison of different rubbers. This includes how different rubbers perform in terms of durability, softness, speed, bounciness, and more.

To complement the unstructured text of the discussion forum, we have also scraped a list of rubbers and overall ratings from revspin.net as groundtruth and to supplement the discussion data when comparison pairs have few discussions. The list includes all the 3,000 possible rubbers currently in production and their ratings in terms of speed, tackiness, price point and ease of use, on a scale of 1 (worst) to 10 (best).

2.2 Data Wrangling & Labeling

To carry out supervised learning, we would need to establish a labeled dataset for groundtruth. Since our forum posts are not all comparative in nature, we have narrowed down to firstly a subset of comparative text by examining the titles using regular expressions. By checking occurrences of "vs", "or", "and", "/", we are able to extract 2,788 comments as the initial dataset.

The second part of the labeling process is by distinguishing comparative sentences out of all sentences.

2.3 Problem Formulation

For any sentence in the comparative comments, it can fall into any of the 5 categories:
0: irrelevant
1: comparative

2: comparative but need context
3: non-comparative statement
4: non-comparative statement but need context

By "need context" means one would need the previous context to know the subject. To represent the comparative relation, it is in the format of [Subject A, Subject B, Attribute, Direction]. Direction is either “>”,”=”.

In the labeling, it is expressed as a list of lists of 4 strings, e.g. [[‘A’, ‘B’, ‘hardness’, ‘>’]].

There can be scenarios where multiple items can be compared in a sentence. For example, A is harder than B and C. Then the relationship will be expressed as two substrings, i.e. [[‘A’, ‘B’, ‘hardness’, ‘>’],[‘A’, ‘C’, ‘hardness’, ‘>’]].

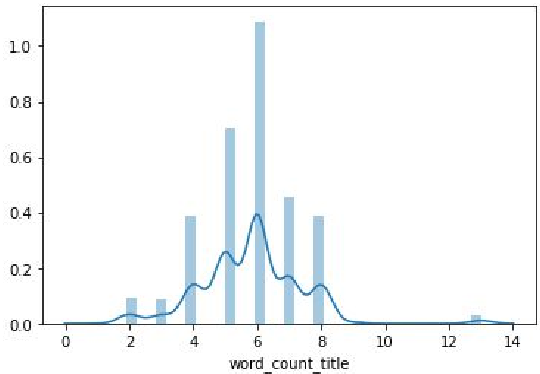
The hierarchical relation expressed through superlative adjective words becomes the direction of comparison. In addition, attributes are expressed as nouns. An example of translating “better” can be seen in this example:
"A is better than B for pushing". The label would be [‘A’, ‘B’, ‘pushing’, ‘>’] (so "better" or "easier" etc. becomes >).

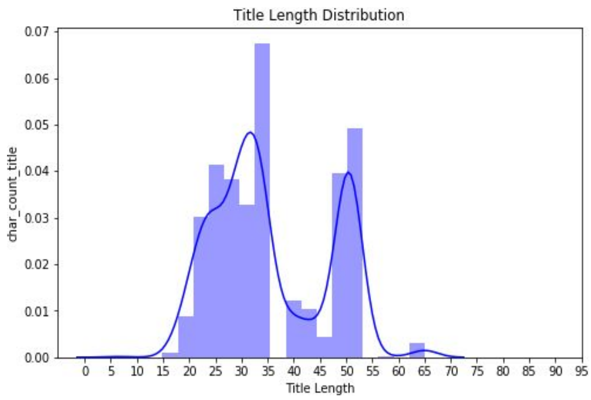
The options for the fourth argument are '>' or '='.

After labeling the data, we then begin to develop our classification model.

2.4 Exploratory Data Analysis

Since our dataset started off as completely unlabelled, EDA is helpful in understanding the text data and to help decide on or develop approaches to labeling and to analyze the data. The below pictures show the distributions of word count and title length.

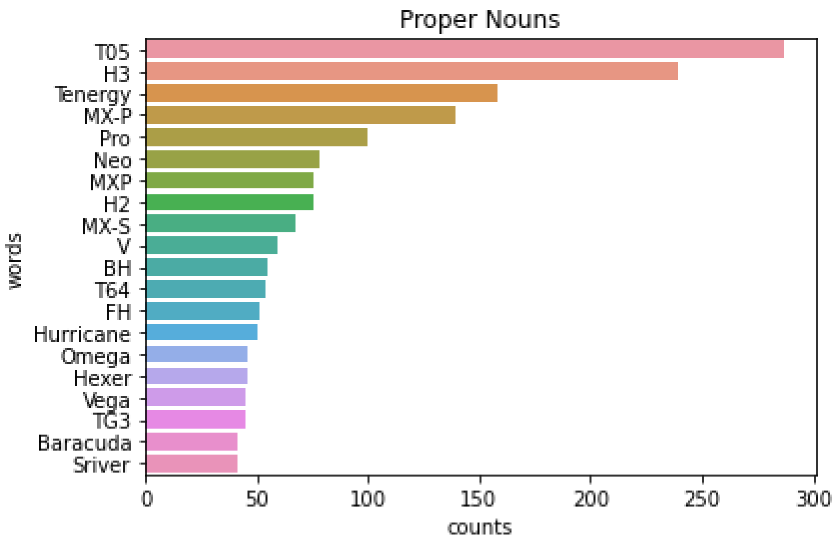




2.4.1 Noun / Adjective Distinction

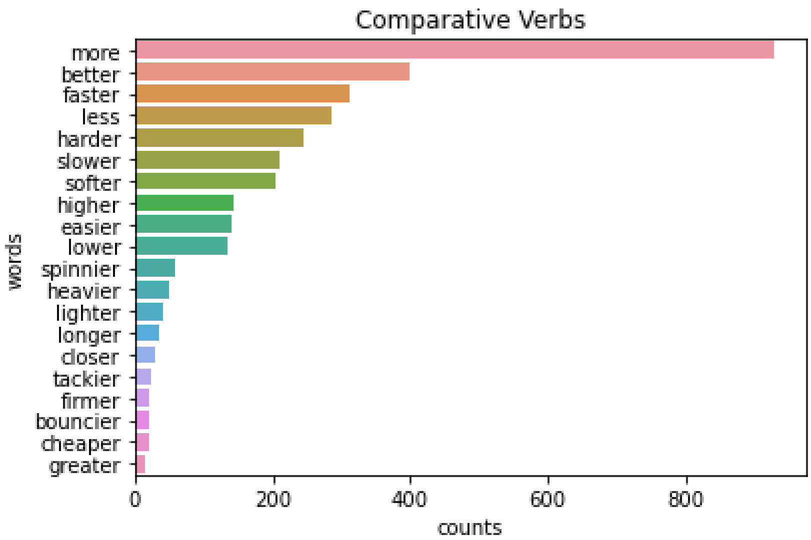
After obtaining comparative sentences assignment from our RoBERTa model, we then perform further EDA on this dataset. Our ultimate goal is to label the two rubbers mentioned, the aspect of comparison in discussion, and find out whether rubber A is better, equal to, or worse than rubber B in that particular aspect / attribute (speed, hardness, etc.). NLTK’s part-of-speech tagging was useful in informing the most proper rubbers and the most popular comparative words in comparative sentences.

The figure below showed some of the most popular rubber names, and the most popular common rubber brands. Additionally, it also helps us discover popular variations of the rubber names, e.g., MX-P and MXP. This is helpful for the entity tagging during model development process.



The most frequent proper nouns in all predicted comparative sentences: T05, H3, MX-P, Neo, H2, MX-S are some of the most common rubbers, with the most popular noun, T05, mentioned approximately 280 times. Meanwhile, Tenergy, Hurricane, Omega, Hexer, Vega, Barracuda are some of the most common rubber brand names, with Tenergy mentioned approximately 160 times. The proper noun identification is done using NLTK POS tagging, filtering only for those with NNP (proper noun, singular) tag.

The figure below shows some of the most frequently used comparative verbs/ adjectives in comparative sentences. This not only helped us in creating the lexicon-based rules to label the comparative words, but also further helped us in identifying the aspects of comparison.



The most frequent comparative verbs/adjectives in all predicted comparative sentences. The top adjectives are either “more”, “less”, or referring to specific features of rubbers, including weight (“heavier”/“lighter”), speed (“faster”/“spinnier”/”slower”) and feel (“softer”/”firmer”). The comparative verbs/adjectives identification is done using NLTK POS tagging, filtering only for those with RBR comparative adverb tags or JJR comparative adjective tags.

2.4.2 Comparative Word Clouds

To help visualize words that are especially prevalent in comparative sentences, we also generated word clouds using the WordCloud module. The figures below show a comparison between a word cloud generated from discussion threads discussing Tenergy (a popular rubber brand), compared to a word cloud generated from discussion threads discussing Barracuda (another popular rubber brand). These word clouds serve as a sanity check for us and are helpful in figuring out whether the rubber brands are particularly different in certain aspects of comparison. As seen in Figure 3, Tenergy generally is considered to have good spin and hard rubber while Baracuda is considered to be faster and better for looping (a table tennis technique).

Word clouds generated using Python WordCloud module using a custom image mask and a custom color scheme. The left word cloud shows the most frequent words for sentences mentioning Tenergy brand rubber, compared to the right word cloud, which shows the most frequent words for sentences mentioning Barracuda brand rubber.

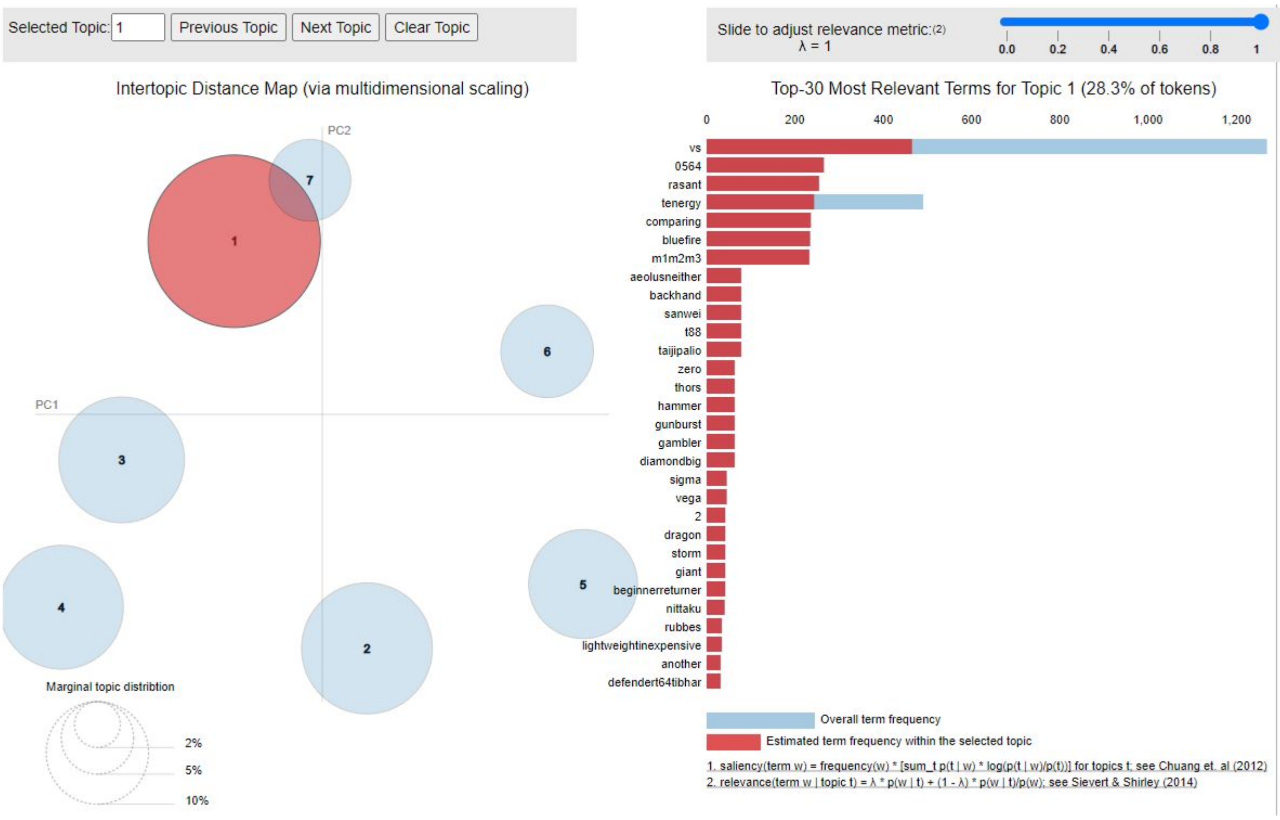
2.4.3 Topic Modeling with LDA

We are using topic modeling which enables consumers to quickly extract the key topics covered by the reviews without having to go through all of them and help the sellers/retailers get consumer feedback in the form of topics (extracted from the consumer reviews). Topic Modeling is a process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus. Topic Models are very useful for multiple purposes, including:

- Document clustering
- Organizing large blocks of textual data
- Information retrieval from unstructured text
- Feature selection

Our aim here is to extract a certain number of groups of important words from the reviews. These groups of words are basically the topics which would help in ascertaining what the consumers are actually talking about in the discussions.

To visualize our topics in a 2-dimensional space, we used the pyLDAvis library. This visualization is interactive in nature and displays topics along with the most relevant words. As we can see in the below picture, there are 7 topics and each topic has the top 30 most relevant words.



3. Statistical Analysis and Machine Learning

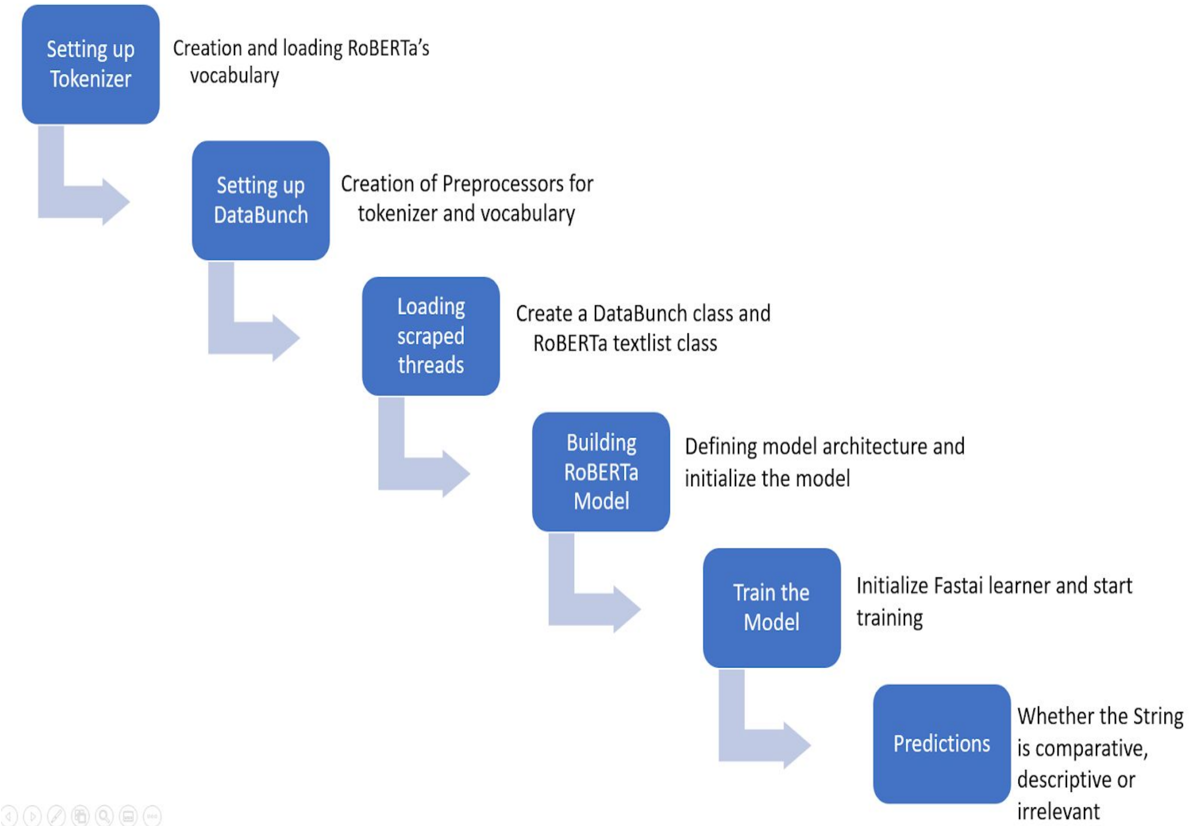
3.1 Comparative Sentence Identification using Transfer Learning

Classification using RoBERTa (Robustly optimized BERT's pretraining approach)

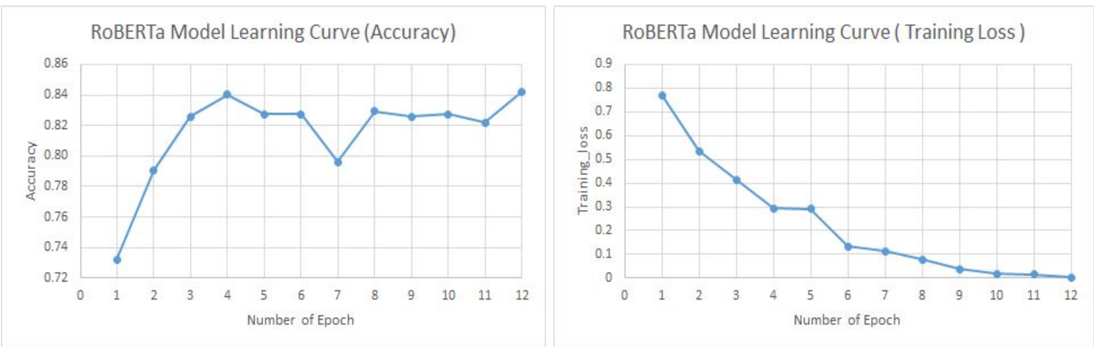
To classify text, we have used RoBERTa (Robustly optimized BERT's pretraining approach) modeling, which gives state of the art results. RoBERTa builds on BERT's language masking strategy and modifies key hyperparameters in BERT, including removing BERT's next-sentence pretraining objective and training with much larger mini-batches as well as learning rates. RoBERTa was also trained on more data in magnitude than BERT, for a longer amount of time. This allows RoBERTa representations to generalize even better to downstream tasks compared to BERT.

Along with RoBERTa we have used the Fastai library, a deep learning library which provides practitioners with high-level components that can quickly and easily provide state-of-the-art results in standard deep learning domains, and provides researchers with low-level components that can be mixed and matched to build new approaches. Fastai provides a streamlined interface to build datasets and train models.

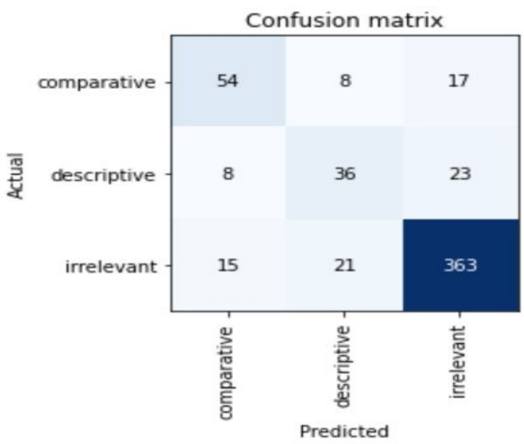
Integrating RoBERTa into Fastai allows us to enjoy the convenience of Fastai methods in combination with the strong predictive power of these pretrained models. Below is the workflow of showing the flow of classification with RoBERTa.



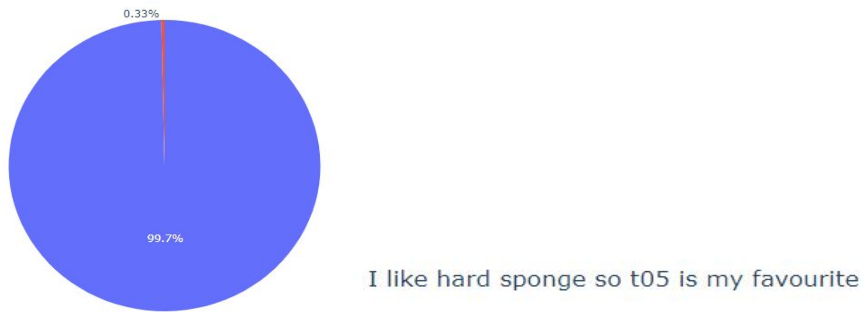
We have trained RoBERTa model on 12 Epochs in which accuracy is increasing and training loss is decreasing after each epoch. The below plots show the RoBERTa model learning curve of accuracy and training loss.



After training RoBERTa model, we are able to classify sentences whether they are comparative, descriptive or irrelevant with 85% accuracy. The confusion matrix shown below for all classes. A confusion matrix is a tabular way of visualizing the performance of your prediction model. Each entry in a confusion matrix denotes the number of predictions made by the model where it classified the classes correctly or incorrectly. As we can see in matrix, there are three classes: comparative, descriptive and irrelevant with predicted values and actual values. Overall, out of all the 545 labelled sentences, most sentences are predicted in the correct category.



After training the RoBERTa model we get predictions as shown below: If we give to the model the string “I like hard sponge so t05 is my favourite”, the classification prediction is comparative with a scale of 99.67% out of 100%.




```
The review class is:  comparative , with a scale of  99.67  out of 100.
```

3.2 Comparison Aspect Detection Using Rule-based Approach

After all sentences in our dataset were assigned as comparative, descriptive, or irrelevant (not comparative) by the RoBERTA model, we then focused on the comparative sentences. From EDA and k-means clustering, we have generated a list of words important for each aspect: speed / spin, hardness, tackiness, control, throw, durability, weight, blocking, looping, chopping, and overall. For example, important words useful in identifying that the users are talking about ‘throw’ are: arc, angle, throw, flat. The presence of any of these words in a sentence helps identify the sentence as talking about ‘throw’. This approach only assigns sentences that contain the keywords used and we then inspect the assigned and non-assigned sentences to read just the keywords we used. We used this naive keywords-based approach because the dataset is unlabeled, and a rule-based approach seems to be a good approach for a first step for proof-of-concept.

After identifying the aspect of comparison, we then want to identify the comparative words in order to assign the direction of comparison (greater, less, equal to). We again used insights gained from EDA and k-mean clusterings to search for the occurrence of specific words like ‘faster, more spin, spinner’ to indicate a greater relationship for speed, and words like ‘not as fast as, about equal in speed’ to indicate an equal relationship for speed, and words like ‘much slower, less speedy’ to indicate an lesser relationship for speed.

We performed manual inspection of the assigned sentences and out of the 30 random sentences we inspected, 23 were correctly identified in a comparative sense, with 7 tagged irrelevant sentences that used the keywords in other contexts. One reason is the RoBERTA model often would assign a question asking about a certain aspect as ‘comparative’ and our keyword-based approach often captures that. Another reason is sometimes people discuss techniques to improve speed. looping, etc, when talking about the rubber even though they are not comparing the rubbers.

One sentence can also contain multiple aspects of comparison but for the purpose of simplicity, we first only focus on sentences with only one aspect of comparison for the next step.

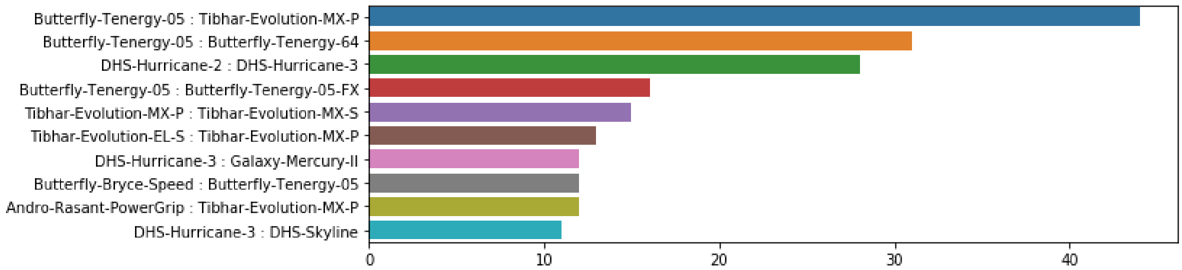
3.3 Entity Identification Using a Lexicon-based Approach

To identify the relevant rubbers in each comment, we use a lexicon-based approach. To build the lexicon, we start from data scraped from RevSpin which consists of a list of almost all table tennis rubbers on the market. For each rubber, we generate a list of possible abbreviations based on common brand and rubber name patterns. For example, the rubber `'Tibhar Evolution EL-P'` could be referred colloquially to as any of `{'Evolution EL-P', 'EL-P', 'ELP', 'Evolution ELP', 'Tibhar Evolution EL-P', 'Tibhar EL-P', 'Tibhar ELP', 'Tibhar Evolution ELP'}`. Based on this dictionary, we map all occurrences of these words in each comment to their respective corresponding rubbers, and identify those as the entities of interest in each comment.

3.4 Determining Comparison Relation Using a Rule-based Approach

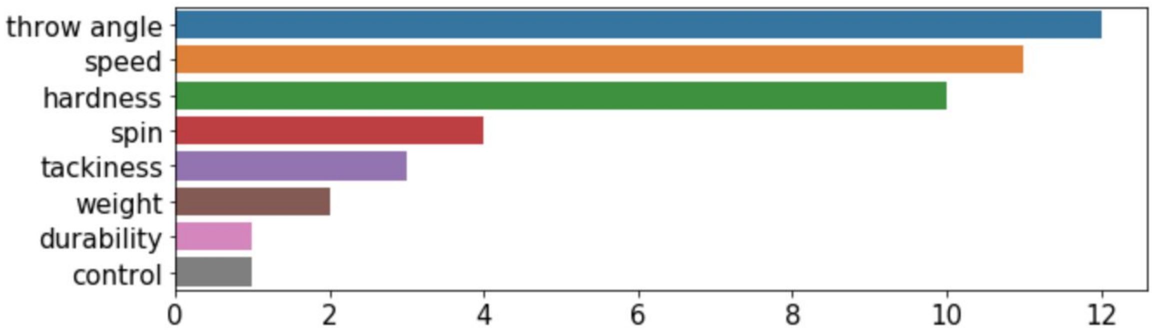
Having identified the entities, aspect of comparison, and comparative word for a comment, we then synthesize them together to form a comparative relation identification. Within a sentence, the immediate predecessor of the comparative word is taken to be the subject of the sentence. For example, in the case ‘Evolution ELP is faster than T05’, ‘Evolution ELP’ would be the subject. For comments where the subject is not explicitly mentioned, for example ‘I think it is faster than T05’, we resort to the thread title to extract the other entity of interest, for example in this case could be ‘EL-P vs T05’. Synthesizing this together, we obtain a relation: [Speed: Butterfly Tenergy 05 < Tibhar Evolution EL-P]. We store the pairs in alphabetical order for data retrieval from the front end.

We can analyze the pairs of rubbers that are being compared. For example, the following figure shows the most frequent comparisons out of all the 100,000 sentences we labelled.



The top two comparisons are Butterfly Tenergy 05 against Tibhar Evolution MX-P and Butterfly Tenergy 05 against Butterfly Tenergy 64. It shows that intra-brand differentiation is something that people care a lot about on the discussion forum.

Of the most frequently compared pair, Butterfly Tenergy 05 and Tibhar Evolution MX-P, the most compared aspects are as follows.



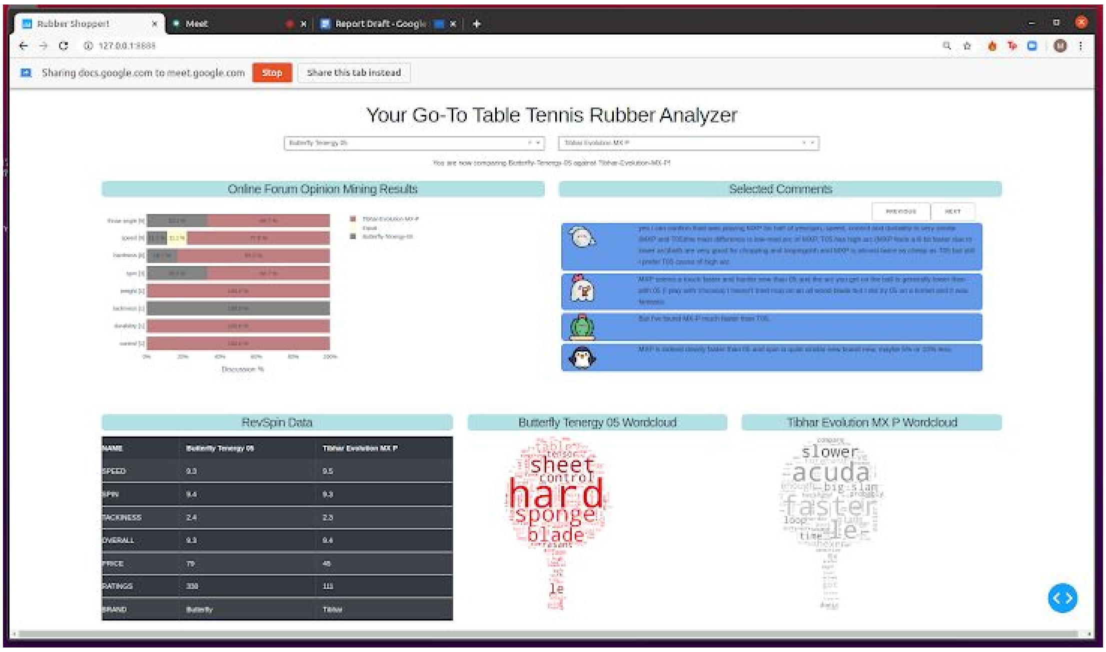
For Butterfly Tenergy 05 and Tibhar Evolution MX-P, the two most compared aspects are throw angle and spped.

4. Front-end Interactive Dashboard

After identifying the comparative sentences and retrieving information concerning the comparative direction, attributes, and entities involved, we have aggregated the results, stored them in an SQLite database and presented the information in an interactive dashboard that runs in Django and DASH frameworks. We have successfully deployed the dashboard with CI/CD integration using Heroku.

In the dashboard, users can choose two rubbers of their preference to compare. After choosing, the dashboard will display several components:

1. Stacked Bar Chart (top left): the chart displays aggregated comparative results by aspect and direction.
2. Display of Selected Comments: when hovering over a stacked bar chart, selected comments will be displayed corresponding to the specific aspect and direction.
3. Quantitative Data from other sources (bottom left): the tabular result is a direct comparison of quantitative data collected from RevSpin.net.
4. Comparative Word Clouds (bottom right): word clouds that display words from comments related to the particular rubber are put side by side to create direct contrast.



Interactive dashboard that integrates all of the model results

All in all, our interactive dashboard allows users to consume the information easily and help them to make informed decisions.

5. Conclusions and Future Work

5.1 Our Findings

Comparative opinion mining is a relatively underexplored territory in terms of the e-commerce space. We decided to specifically focus on performing comparative opinion mining on table tennis rubber due to our painful experiences with trying to compare two rubbers. We completed an end-to-end project within 5 weeks and created a live interactive dashboard to help consumers make informed decisions when comparing two rubbers.

We began with collecting more than 16,000 threads from table tennis forums, performed EDA, data cleaning and manually labeling some comparative sentences. We then trained a transfer learning RoBERTa model to 84% accuracy to help predict comparative sentences. Next, we created lexicon-based rules informed by EDA and literature to perform entity identification and aspect comparison. Lastly, we built a Dash dashboard frontend with a SQLite backend to provide an interactive dashboard for users to compare and visualize two rubbers in terms of different aspects. This dashboard was then deployed using Heroku.

Our findings would be very helpful for both customers and suppliers of table tennis rubbers. The project is highly applicable, reusable and scalable, as we analyzed data from top quality and popular rubber products, scraped from the most commonly used review websites for table tennis enthusiasts.

Regarding consumers, the interactive dashboard would provide a more effective and enjoyable shopping experience, which involves cross-comparing different rubber products for features of preference, including ease of use, firmness, and durability. This could further develop into a sophisticated central platform for product review analysis and recommendations among table tennis communities. The benefits include great time savings, more active product reviewing, and better product choices.

Regarding suppliers, the rich information on their product performance in comparison with competitors' would inform more effective and targeted manufacturing, marketing, customer service, and other key aspects of the business. The benefits may include cost savings, revenue growth, boosted customer engagement and loyalty. In the long run, data-informed strategies will create competitive advantages for suppliers and high quality products for consumers.

5.2 Future Work

Time permitting, we would like to collect a larger dataset by scraping more consumer reviews on table tennis rubbers. We would analyze all threads in the dataset for comparative opinions, instead of threads with comparative titles alone. For the purpose of the project, we are using a rule-based approach to identifying aspects of comparisons; however, it would be interesting to implement machine learning models for this purpose. With more data, it would be possible to form relationship graphs, and use transitive relations to generate more comparative pairs. For future work, it would also be exciting to have large and representative datasets of various consumer products to extend from our work on table tennis rubber reviews. It would also be interesting to explore the interconnection between rubbers using graphical models. We encourage interested researchers to expand and improve upon our work so that consumers and suppliers of various products and services can benefit from effective comparative opinion mining.

References

Xu, Kaiquan, et al. "Mining comparative opinions from customer reviews for competitive intelligence." *Decision support systems* 50.4 (2011): 743-754.

Varathan, Kasturi Dewi, Anastasia Giachanou, and Fabio Crestani. "Comparative opinion mining: a review." *Journal of the Association for Information Science and Technology* 68.4 (2017): 811-829.

Ganapathibhotla, Murthy, and Bing Liu. "Mining opinions in comparative sentences." *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. 2008.

Jindal, Nitin, and Bing Liu. "Identifying comparative sentences in text documents." *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006.