Amazon Product Review Visualization

Aditya Kanchivakam Ananth - akanchiv@asu.edu
Deepan Karthik Nedumaran - dnedumar@asu.edu
Gowtham Sekkilar - gsekkila@asu.edu
Motthilal Baskaran - mbaskar2@asu.edu
Srivathsan Bhaskaran - slbhask1@asu.edu

Abstract— **The deluge of e-commerce sites enables people an easy way of shopping at any time and from anywhere in the world to their desired place. People often order online whenever they need something and are driven through their intellect and expertise in choosing their desired product. However, when it comes to children who are being voided of those privileges, the parents play a vital role from choosing the quality to ordering from online, in most cases by making use of the reviews of the products. How easy and helpful are these reviews to the customers, especially in our case where the eldest have very little idea of what works well for kids, demands answering.**

To resolve the above-mentioned problem, we implement a dashboard that reveals the product review analysis for a product chosen of how the review stands out from others and really help in adding value to the product.

Keywords – Sentiment Analysis,

I. INTRODUCTION

Today, Amazon and Yelp lead the market in the field of e-commerce with its huge variety of products in almost every field and the reviews have turned out into widely used references for decision making. However, there exists a constraint on them as there might be huge amounts of them and in unorganized fashion as well. This might worry the customer a bit as they might feel them to be profusely long and could end up skimming the content. The whole purpose of writing a review gets lost in such occasions. Amazon is the most widely used e-commerce site for product purchase making use of the reviews, ratings, and comments about the products. In our case, it is highly likely that parents make use of these reviews in making the order placing decisions. There is a whole lot of data available in this section that is not readily intuitive to the customer making the easy task cumbersome.

In our project, we have chosen the Amazon customer review dataset from the Amazon website available online. In our cause, we have selected the “Toys” dataset counting more than 2M with almost 500+ products and found a lot of data are skewed. Hence we try to filter out the top 100 products based on the reviews. With this as the base data, we build our analytical dashboard that gives customers an idea about the review overview, sentiment, quality, distribution, most frequent patterns and also category (content) based recommendation with Product ID as input.

We have used JavaScript, D3.js for interactive visualizations, Python code for sentiment analysis, helpfulness determination and for various data operations like cleansing, pre-processing, adding additional calculated columns needed for visualization, transforming existing data format into required one, etc. Since we deal with top product review data, each source file have been moved to the server and are fetched upon request from the respective script. Several principles that were taught in class have been incorporated thereby rendering an elegant, coherent and intuitive visualization that makes the customer’s task easy and gives most impacting data. The below sections throws a spotlight on the different project aspects like project overview, components, way of implementation of the dashboard, my contribution towards the project, what I have learned from it, the results obtained and the evaluations observed.

II. MOTIVATION

For a customer trying to understand the worthiness of a product, analyzing reviews would be of great help. But too many reviews to sort through can be overwhelming. Providing coherent insights visually, would help them to make a decision easily quickly. Especially for children a comprehensive analysis of the reviews of a product would be helpful for a better understanding and decision making. Hence a Dash Board is formed which displays the insights about the reviews of the product through various visualizations and this is helping the user analyze the product through the merits of the review. The Dash Board is a simplified version of the reviews visualised through Line and Bar chart, Gauge Chart, Word Cloud. The Product title, price of the product, Total number of reviews, Sales Rank aide in the decision making process. Thereby making the purchase decision making a breeze.

III. DATA SET

The Amazon product reviews data set from SNAP was processed and used for the visualizations. The data span a period of 18 years, reviews up to July 2014. This data set includes reviews (ratings, text, helpfulness votes), and product meta data (descriptions, category information, price, brand, and image features). The review data has the following information:

- reviewerID - ID of the reviewer
- asin - ID of the product assigned by Amazon
- reviewerName - name of the reviewer

- helpfulness - helpfulness rating of the review, e.g. 2/3
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- reviewTime - time of the review (raw)

We have imported some information from the meta data file and incorporated it to our data set. The meta data includes the following information:

- title - name of the product
- price - price in US dollars (at time of crawl)
- related - related products (also bought, also viewed, bought together, bought after viewing)
- categories - list of categories the product belongs to
- Sales Rank - The selling rank of the particular product in the respective category
- related - related products (also bought, also viewed, bought together, buy after viewing)
- categories - list of categories the product belongs to.

"Toys" data-set has been selected from the SNAP for counting more than 2M with almost 500+ products and found a lot of data are skewed. Hence we try to filter out the top 100 products based on the reviews. Using this as the base data, analytical dashboard is built that give users an idea about the review overview, sentiment, quality, distribution, most frequent patterns and also category (content) based recommendation with Product ID as input.

The data was in the format of JSON, which was converted to CSV format for pre-processing as CSV is easier to transform and edit to the required formats. The main CSV file has the following topics: toys-asin, toys-helpful-0, toys-helpful-1, Helpfuness, toys-overall, toys-reviewText, toys-reviewTime, toys-reviewerID, toys-reviewerName, toys-summary, toys-unixReviewTime, Date-int, Date, Year. These are obtained after pre-processing the data which has been taken from SNAP. Date which was mentioned in the Unix format has been converted to standard time for convenience. Along with the mentioned data, the meta data was added for further data creation for the visuals.

A. Visual Data Files

1) Rating over the years: The file has the rating data for the products which were reviewed by all the users who bought the product. The data has been presented in the form of line graph which is a representation of time series. The value presented is the weighted average of the ratings for the year. it is Asin and Ratings and year.

2) Ratings for the given year: It is the count of the each of the ratings for the year for the particular product. It has the data Asin, Ratings and year. This is visualised through a bar chart

3) Sentimental Gauge: The file has the positive and negative ratings for the respective Asin ID's. This represents how much percentage of the reviews are positive reviews and negative previews. This data is presented using gauge chart which is represnted by the chart with green color as the positive percentage and the red color gauge as the negative percentage

4) Word Cloud Gauge: The file contains the Asin ID and then reviews for each of the product. The Reviews are passed through NLTK parser and the data for the word cloud is then passed through zing charts for visualization. This helps in analyzing the most used words in the reviews. Just a look at the word cloud should help the user in seeing the top used words throughout the product review.

5) Product-Categories: This file contains the data for asin ID and the categories respectively. Another file contains the Categories and the top recommended products in each of the categories. This is useful for recommending products belonging to the similar categories. This gives the user a choice of choosing similar products as they all fall into the same category

6) Helpfulness Gauge: The file contains the value of Asin ID and the helpfulness percentage for the respective reviews. This data is visualised through HighCharts. This helps us analyze how helpful the reviews have been till now. This helps the user to make a better judgement in deciding about the product

IV. Visualization Design

Using charts or graphs to visualize large amounts of complex data is easier than pouring over spreadsheets or reports.

A. Line Chart

The Line chart is represented by a series of data points connected with a straight line. Line charts are more often used to visualize trends in data over time, hence the line is drawn in a chronological . Data points are plotted and connected by a line in a "dot-to-dot" fashion.

B. Bar Chart

The classic Bar Chart uses either horizontal or vertical bars (column chart) to show discrete, numerical comparisons across categories. One axis of the chart shows the specific categories being compared and the other axis represents a discrete value scale. Bar charts are distinguished from Histograms, as they do not display continuous developments over an interval.

C. Packed Bubble Chart

A bubble chart requires three dimensions of data, the x-value, and the y-value to position the bubble along the value axes and a third value for its volume (size). Packed Bubble charts have a simpler data structure, a flat, one-dimensional array with volumes. The bubble's x/y position is automatically calculated using an algorithm that packs the bubbles in a cluster.

D. Solid gauge

Solid Gauge Chart is similar to the Angular Gauge Chart and is most commonly used to mimic real-world gauges. The main difference from the Angular Gauge Chart is that the values are displayed by a filled portion of a gauge scale rather than a hand of a mechanical-like gauge.

E. Word Cloud

A word cloud is a visualization of a set of words, where the size and placement of a word are determined by how it is weighted. The more a specific word appears in a source of textual data, the bigger and bolder it appears in the word cloud.

V. Methodology

The Dashboard bestows a simple, elegant and effective visualization for customers by providing something beyond than what can be interpreted by reading a review directly. It does indeed, add an additional perspective to the viewer by enabling them to easily grasp what the visualization tries to convey about the product review.

A. Rating Analysis

1) Average Ratings vs Year- Line Chart: The graph is visualized as a Line chart since they are best used to display quantitative values over a continuous interval or time period. A Lot of reviewers give ratings for each product every year. The average of all the ratings in each year is used to plot the points in the line chart. Interactive Line Chart shows the average ratings over time that gives an idea of how the review trend of the product has changed over the years. The ups and downs in the graph can be attributed to changes in the product sales and its impact on customers. When hovered over a particular year in the chart, the rating distribution for that year is displayed. On clicking the year, a bar chart pops up with the rating distribution.In here, the number of ratings ranging from lowest to highest can be interpreted.The color schemes have been chosen so that the readings are easily seen to the customer and are also relevant to the Amazon products.

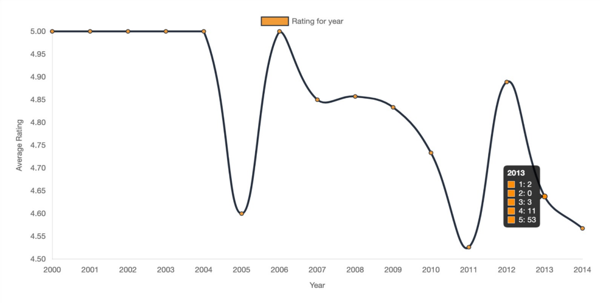


Fig. 1. Rating over the years

2) Rating Distribution in a particular year: In the Line chart, when clicked on a particular year trend, it pops up a bar chart that gives the relation between the rating of that particular year versus the rating frequency. The charts also comes with hover capability that shows the rating split which gives supplementary information to the customer that could save extra effort if the user feels that information is sufficient.

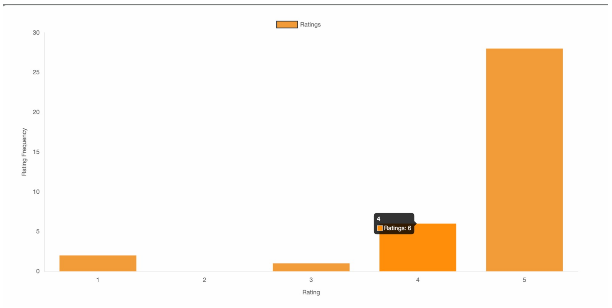


Fig. 2. Ratings for the given year

B. Sentiment Analysis

1) Positive score and Negative score: The gauge shows the positive and the negative score percentage calculated based on the review comments of the product. Higher the positive percentage, the better is the review of the product which bolsters the product purchase possibility. Higher negative percentage score tells that the product is not liked by a greater part of the customers who had purchased them already from the amazon market. The use of universal colors green and red to represent positive and negative scores make the visualization more intuitive to the viewer.

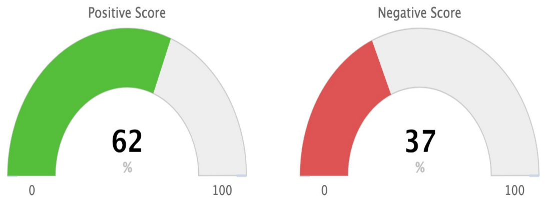


Fig. 3. Sentimental Gauge

2) Word Cloud: The next visual is the review based word cloud visualization. The word cloud displays the most frequently used words in both positive and negative aspects and the user can view the frequency count by hovering the particular word. This drills down the sentiment score gauge by displaying the most frequently used words in both positive and negative aspects and the user can view the frequency count by hovering the particular word. This word cloud extends the former view rather than just revealing the essentials. It provides an emotional connection and gives the feel of engaging with the process flow with minimal use of colors. The sequence of using word cloud after sentiment gauge has an upper hand in explaining the context to the user. Had word cloud been used in prior to sentiment, the user would have definitely experienced the clumsiness around. Hence this approach once again proves to be the better usage option.

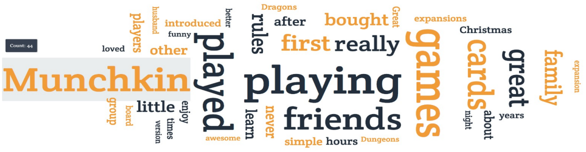


Fig. 4. Word Cloud

C. Helpfulness Gauge

The Helpfulness chart provides the number that actually gives the estimate of how much the reviews help the cause - in helping to recommend the product typed in. The review helpfulness data has been used with some calculation in determining this number. A considerable high value suggests that the reviews have gone through so far indeed might help the customer to buy the desired product. Relatively, this score can be taken into consideration directly while thinking to buy a product.

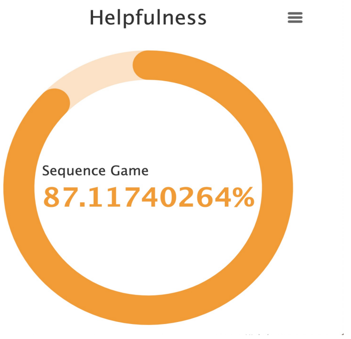


Fig. 5. Helpfulness Chart

D. Recommended products for categories

Category-based product recommendation does the actual recommendation, based on the analysis done on the data set in the form of a bubble chart. The bigger circles correspond to the categories and the smaller ones correspond to the top products in the respective category. For a particular product ID, the toy data set can have up to five categories which further grounds into their relevant products. Also, when hovered on the categories, there is a list of top products being displayed as part of the evaluation process.

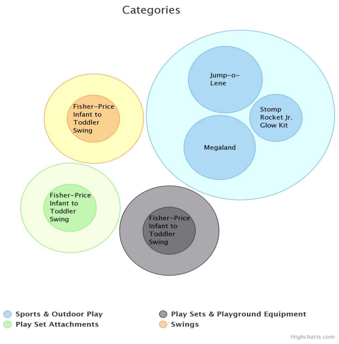


Fig. 6. Category-Product chart

VI. COLOR SCHEME

The Color code used for the project is Amazon color codes which has been

- Blue - RGB (37, 47,62)
- Yellow - RGB (242, 156, 56)
- Grey - RGB (234, 237, 237)

VII. TECHNOLOGIES USED

- HTML5 - hyper Text Markup Language is a software solution stack that defines the properties and behaviours of web page content by implementing a markup based patterns to it
- JavaScript - A high level interpreted programming language which is used for the interactive web pages and is core for the World Wide Web
- CSS - Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language like HTML
- D3.js - Data Driven Documents is a JavaScript library for producing interactive, dynamic data visualizations in web browsers. It allows control over the final visual result
- Charts.js - A JavsScript library that is used for creating interactive charts which is easier than D3 for creating interactive charts. It has more features embedded into it for visual, interactive and accessible elements.
- HighCharts - SVG based multiplatform chart library that is easy to add interactive, mobile-optimized charts to your web and mobile projects. It features robust documentation, advanced responsiveness
- NLTK - A platform which is used to interact with Human Language Data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries
- JSON - JavaScript Object Notation) is a lightweight data-interchange format. It is easy for machines to parse and generate. JSON is built on two structures:A collection of name/value pairs. In various languages, this is realized as an object, record, struct, dictionary, hash table, keyed list, or associative array and An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence.
- JQuery - jQuery is a JavaScript library designed to simplify HTML DOM tree traversal and manipulation, as well as event handling, CSS animation, and Ajax.
- Ajax - A set of web development techniques using many web technologies on the client side to create asynchronous web applications. With Ajax, web applications can send and retrieve data from a server asynchronously (in the background) without interfering with the display and behavior of the existing page
- Python - Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python features a comprehensive standard library, and is referred to as "batteries included". This is used for the data pre-processing
- Zing charts - A JavaScript API which has been used for the creation of wordclouds. This is a set of modules which has been used for creating the interactive, dynamic and responsive mobile charts for Big Data

VIII. Web Hosting

Amazon Web Services has been used for hosting the webpage. EC2 - T2.micro is the instance which has been used for the project. The use cases of T2 Instance are for Websites and web applications, development environments, build servers, code repositories, micro services, test and staging environments, and line of business applications.

T2 instances are Burstable Performance Instances that provide a baseline level of CPU performance with the ability to burst above the baseline.

T2 Unlimited instances can sustain high CPU performance for as long as a workload needs it. For most general-purpose workloads, T2 Unlimited instances will provide ample performance without any additional charges. If the instance needs to run at higher CPU utilization for a prolonged period, it can also do so at a flat additional charge of 5 cents per vCPU-hour.

The baseline performance and ability to burst are governed by CPU Credits. T2 instances receive CPU Credits continuously at a set rate depending on the instance size, accumulating CPU Credits when they are idle, and consuming CPU credits when they are active. T2 instances are a good choice for a variety of general-purpose workloads including micro-services, low-latency interactive applications, small and medium databases, virtual desktops, development, build and stage environments, code repositories, and product prototypes. For more information see Burstable Performance Instances.

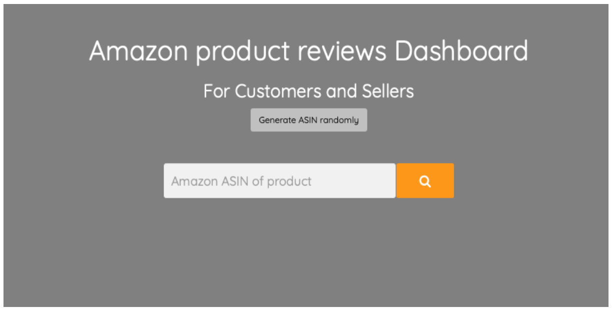
Features:

- High frequency Intel Xeon processors
- Burstable CPU, governed by CPU Credits, and consistent baseline performance
- Lowest-cost general purpose instance type, and Free Tier eligible*
- Balance of compute, memory, and network resources

T2.micro has 1 vCPU, 6 CPU credits/hour, 1GB Memory and has a low to moderate Network Performance.
The Website link hosted using the EC2 T2.micro server is http://ec2-3-16-24-222.us-east-2.compute.amazonaws.com:8000/

IX. Evaluation Plan

The team started off with the "Field study" gathering insights and ideas about different e-commerce websites and dashboards, collecting ideas and techniques to be incorporated that match the project requirement. The good and the bad ideas used in those references were noted and best of those are compiled to be implemented in our dashboard. This was implemented as the using the web-page format as the below image



Amazon product reviews Dashboard

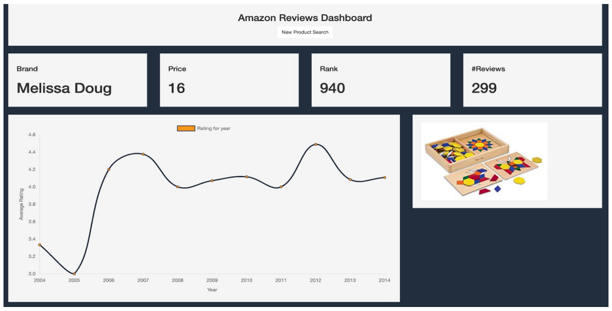
For Customers and Sellers

Fig. 7. Review Home Page

This was used as the dashboard proved to be more useful and intuitive for the reviews project and this made the presentation of data in a dashboard format which presents the data in a related fashion and helps the user understand the product through the various visualisations.

Next step was the Cognitive walk-through carried out by each team members. Insights gathered from the field study were made into prototype on paper and were cross inspected by the remaining team members to find the defects and rectifications were done. After going through the created prototypes, the best ones were picked and were implemented in the final dashboard.

The Final presentation of the dashboard is as follows:



Amazon Reviews Dashboard

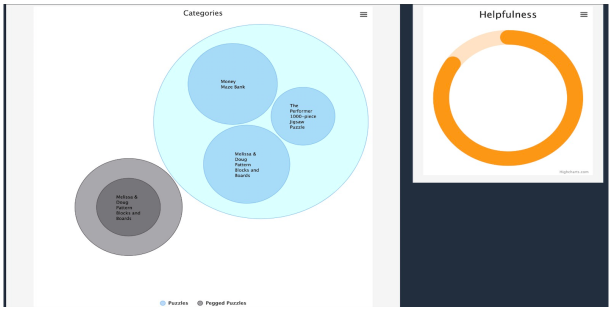
Brand: Melissa Doug

Price: 16

Rank: 940

#Reviews: 299

Fig. 8. Dash Board 1



Categories
Money Maze Bank
The Performer 1000-piece Jigsaw Puzzle
Melissa & Doug Pattern Blocks and Boards
Melissa & Doug Pattern Blocks and Boards
Puzzles
Pegged Puzzles
Helpfulness
Highcharts.com

Fig. 9. Dash Board 2

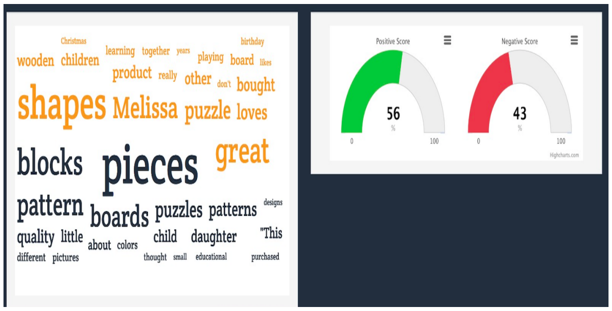


Fig. 10. Dash Board 3

Last step in deciding the final format of the dashboard was the "Heuristic Evaluation." The created dashboard were tested with the general audience from all age groups who use the e-commerce websites. The Heuristic evaluation score feedback from the audience were considered and several iterations were carried out with fixes in each stage until the final version was reached that satisfied nearly 98% of the whole audience involved it the whole process. Bug fixes were done in each iterations, enhancing the user experience with the dashboard with usability testing done in parallel. The final version of the dashboard is released after through round of validations with the vision of the project met.

The response from the users are as follows:

- "Review based frequency word cloud was unique. Instead of regular word cloud, they pre-processed the text and then created it which made it more informative"
- "They analyzed data from amazon. I think it's pretty useful to provide the most frequent work appearing in the comments. The implementation is simple and elegant"
- "The color choices and the problem statement is good, the word cloud should have shown the features of the product that has been talked about a lot instead of just text that has been used more often"
- "Amazon Product Review Dashboard view is unique and consolidates the view and the data gets populated based on the input product"

X. Discussion Future Works

The Product recommendation can be improvised with the use of more efficient Machine Learning Algorithms which would provide more accurate recommended products for the user. Improvise better and more efficient Machine Learning algorithms to expand the product recommendation, Natural Language Processing to improve the genuity of the reviews thereby improving the user experience.

XI. ACKNOWLEDGMENT

I would like to thank Professor Doctor Sharon Hsiao for having taught me this course, and also the Teaching Assistant Yancy Vance Paredes for having guided me throughout the project. This project would not have been successful without their constant guidance. I would also like to thank my team members for having put in their contributions and also helped me with my parts which resulted in this project being successful.

XII. CONCLUSIONS

The dashboard created has helped resolve the problems faced by children and parents while ordering toys for children. This has helped people give an insight to all the reviews and the analysis from the reviews has helped people make better decisions. It does indeed, add an additional perspective to the user by enabling them to easily grasp what the visualization tries to convey about the product review. The idea of static and interactive use of visualization from the class assignments was also incorporated in the project that entitles pattern predictions and the attributes that influence it. This can be further expanded to all the categories as the dashboard is based on the product reviews. Thus the Dashboard provides a cluster-free experience for the users

References

[1] Tanjim Ul Haque ; Nudrat Nawal Saber ; Faisal Muhammad Shah.:Sentiment analysis on large scale Amazon product reviews.In: IEEE International Conference on Innovative Research and Development(2018)

[2] Shashank Kumar Chauhan, Anupam Goel, Prafull Goel, Avishkar Chauhan and Mahendra K Gurve.:Research on Product Review Analysis and Spam Review Detection.(2017)

[3] Xing Fang and Justin Zhan.Sentiment analysis using product review data.(2015)

[4] Social Visualization Encouraging Participation in Online Communities (2006) Lingling Sun, Julita Vassileva

[5] Design principles for visual communication. (2011).Agrawala, Maneesh, Li, Wilmot, Berthouzoz, Floraine. Commun. ACM, 54(4), 60-69. doi: 10.1145/1924421.1924439

[6] Evaluating Information Visualizations (2008) Sheelagh Carpendale, Information Visualization, Lecture Notes in Computer Science Volume 4950, 2008, pp 19-45

[7] Pandey, A. V., Rall, K., Satterthwaite, M., Nov, O., Bertini, E. (2015). How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion Techniques.

[8] Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., Carman, M. (2016). Are Word Embedding-based Features Useful for Sarcasm Detection?. arXiv preprint arXiv:1610.00883.

[9] http://jmcauley.ucsd.edu/data/amazon/

[10] https://www.highcharts.com/

[11] https://www.chartjs.org/

[12] http://zingchart.com/