1. Thanks for coming to my presentation today, my name is Ayan Karim and for my final capstone project, I developed an Opinion Mining Product. Which, I’ll explain what that mean in a little bit.
2. So, data scientists we’re already aware of the immense amount of data that companies collect every day. But beyond the company's own databases exist tons of information online that they don't have direct access to. Any company that sells a product in mass production generates a huge amount of public opinion on the product and tons of reviews, articles, reactions and overall sentiments are posted online, especially soon after release.

So now when companies see this wealth of information that isn't a part of their own analytics, they want to access it to learn about how they can make their products better, or how they're products are performing.

The problem is, how do companies access this data from the public media and turn it into something useful? How do we extract all that diverse information from the web and analyze it? This is where Opinion Mining comes in.

1. What is Opinion Mining? "Opinion Mining" is an application of Data Science in which a pipeline is designed to process and interpret public opinion about various products. The source of this data usually comes from public sources like reviews, article and social media, and these text documents allow companies to understand of how people feel about their products.

To do this, we have to do a type of analysis called Aspect-Based Sentiment Analysis. This basically means our model will extract sentiments of a product within context. So, the model will parse the text and extract the sentiment as well as the aspect that the sentiment is about. We’ll see what that means later on in the presentation.

1. So, for the sake of my project, I pretended that I was apple. And as a data scientist for apple, I wanted to inform my company about public sentiments on the iPhone X so that they can use that information for improving their smartphones in the future. Furthermore, I wanted to see how our product compares with the Samsung Galaxy S9, to see the differences between us and our competition.

To guide my investigation, I asked three basic questions.

* What does Public Opinion tell us about the iPhone X?
* What do people dislike about the iPhone X?
* How does our product compare with that of a competing smartphone like the Samsung Galaxy S9?

1. I developed this project in three parts.
   1. First, I had to collect my data by scraping articles about our phones from the web.
   2. Second, I don’t know anything about the context of the texts that I collected, and I wanted to show my results grouped into topics, so that our end-user can clearly interpret the sentiments. So extracted topics from our text documents, to understand the main points that the text talks about.
   3. Finally, I conducted the Aspect-Based Semantic Analysis where I extracted public sentiments about numerous aspects related to the iPhone X.
2. So, the first thing that I needed to do was collect as much data as I could. I scraped three tech websites for text data on each of these phones. As you can see, I collected more data than just for the iPhone or the Galaxy. That’s because I needed as much text data about phones to extract sparse topics.
3. Basically, my scrapers crawl through search queries like this, for each phone, and extracts all the text data in each link.
4. Then the data is magically cleaned placed into a nice data-frame like this.
5. Finally, my product uses Latent Dirichlet Allocation to extract topics about the text.
6. This table kind of shows you a visual representation of the topics, how significant they are, and how well separated they are.
7. Now, once I have the topics, I have to manually, interpret what the topics are based on the most significant words in each topic cluster.
8. I came up with Design, Function and Reliability.
9. Then, the product assigns the labels to the text using a list comprehension algorithm.
10. Finally, a multi-class SVM classifier is trained on the data.

The reason that my product went through the trouble of modeling the topics, is so we can group the sentiments later on, based on these topics.

1. Now we’re ready for the Aspect-Based Sentiment Analysis.
2. We’re only working with our data in question now. We want the sentiments about our product, the iPhone X, and compare it with the Samsung Galaxy S9.
3. This analysis happens in three steps.
   1. It uses an algorithm to extract sentiment words and aspects, sentence by sentence. And uses an opinion lexicon to identify the sentiment attached with the noun. Then, another algorithm tallies the sentiments for each aspect and group them into topics. Finally, we’ll visualize the results for interpretation.
4. Now we have the final output of our results. We have two tables, one for iPhone X and one for the Galaxy, that show the most negative aspects of our phones, grouped into topics.

Now we can interpret these results. From viewing the negative aspects of iPhone X, we see that one of the biggest problems with the phone is how easily it breaks. We see that "scratch", "scratches", "bumps", "damage", "shock", "falls" and "cracks" are shown as the most negative aspects of design and function, almost in sequence. So clearly iPhone X has a problem with its durability. Other problems, such as "cost" is also shown.

However, from viewing the negative aspects of Samsung S9, we see that they have a more diverse set of problems from "video", "lag" to "motion" and "battery". So, it may be encouraging for iPhone X that at least its problems are consistent and can be focused, where as that of the Galaxy S9 seem to range a bit.

1. We want to compare the results of these two phones a little more. So my product produces a table and a bar plot, showing the ratio of negative aspects from the iPhone X to the Galaxy.

When we look at the ratios of sentiments row by row, we can see some concerning results for iPhone X. There's a higher ratio of negative sentiments for iPhone X (1.193) and a lower ratio of positive sentiments (0.897). The same trend is seen for the ratio of sentiments after normalizing both positive and negative sentiments for the total amount of sentiments (neg\_norm: 1.229 and pos\_norm: 0.91)

At first, this could be attributed to class imbalance i.e. there may be more opinions about the iPhone X just because it encompasses a larger market. But that isn't the case because a similar number of documents were analyzed for both phones. Furthermore, if class imbalance were responsible, then there would be more positive sentiments for iPhone X as well.

1. This bar graph kind of shows how the ratio of negative sentiments for the iPhone X is a lot higher than the ratio of positive sentiments.

From this information, we can see that our there's a lot of negative opinion about the iPhone compared to that of the S9. iPhone X should consider doing some serious product development in order to change their public opinion

1. The next steps for this project should include:
   1. Setting up a feature for standardized streaming of data that updates our datasets in intervals.
   2. And developing a recommendation engine to automatically recommend which aspects should are most actionable for product development based on our results, without having to interpret them manually.
2. There’s also a lot room for improving this product. We can always scrape more data, which will improve our Topic and Classifier models.

Also, my algorithms seem to have trouble interpreting bi and trigrams, as well as some very tech specific vocabulary, and I haven’t figured out how to fix this problem. But that should definitely be handled through some research.

1. So that’s all for my presentation today, I hope everyone enjoyed my walkthrough, and hopefully you all see the value in a product like this. I’m more than happy to answer any questions if there are any. Thanks, again.