Ayan Karim Proposal

01/14/2019

**Opinion Mining using Aspect-Based Sentiment Analysis for Analyzing Competition and Product Development**

**Problem**

Tech companies want to assess their products’ performance and their consumer’s opinions to their products. Most companies have their own forum for complaints and methods of collecting feedback, but most people don’t place general reviews and opinions directly on those forums unless they feel strongly about something specific that’s affecting them, so there’s a selection bias.

Furthermore, they want to understand how their product compares with that of their competitors, not just in sales but in performance.

To get the whole picture of how consumers feel about their product as well as that of their competition, the company needs information from public media like news and social media.

The **problem** is that there’s an opportunity cost in the time spent on researching reviews online and deciding which reviews are meaningful to a company for product development, then repeating that research on competitors for the sake of comparison.

**Value**

My product will do the following:

1. Collect consumer opinions on a company’s product and that of their competitor from popular media sources online.
2. A semantic analysis of the reviews and categorize their semantics by positive, negative or neutral.
3. It will cluster the text into topics (memory, speed, user-friendliness, reliability, etc.) that are related to positive and negative sentiments and sort the aspects (terms that refer to one of the topics) under each sentiment, based on popularity.
4. The results will return two queries, one on your companies’ opinions and one on your competitors.
5. Finally, my product will produce visualizations that compare the two companies’ opinion data based on their positive and negative sentiments and how they differ in each topic.

The **value** that my product gives is return an easy to read and visualize query of popular opinions on a company’s product as well as which aspects the opinion refers to. And it compares these opinions to that of their main competitor. A company can then use this information for product development.

**Data Source**

The data will be textual information scraped from TechRadar, Wired and Twitter.

**Techniques that will be used in this project**

* Natural Language Processing
  + Spacy for text processing
  + Sense2vec and TF-IDF Vectorizers for feature engineering.
  + LSA and its variants for topic modelling.
  + TextRank for popularity and significance
* Linear Support Vector Classifiers
  + Classification approach to topic modelling.
* Logistic Regression
  + May use a pre-trained model to use classify sentiments
* Neural Network
  + Neural Coeff v2.0 for pretrained neural networks that parse and replaces pronouns

**Possible Challenges**

* My biggest challenge will be scraping the data and cleaning it into a form I can process and explore.
* Figuring out how to analyze the semantics of textual data that don’t have labels. I’ll probably use a pretrained model.
* Producing the resulting tables for each company so that they’re easy to compare side by side and produce the accompanying visualizations.