Gaining Business Value From Unstructured Data

Why and how should we extract insight from unstructured data

SHARE THIS POST:

September 14, 2016
by Marina Jeon

With the explosion of big data, there has been a reciprocal explosion of companies trying to mine value from the overwhelming amount of data out there. When looking at the data that needs to be analyzed, we can find two distinct types: structured and unstructured data.
We are all aware of structured data: purchases, transactions and electronic sign-ups, but…what is unstructured data? How is it different from structured data?

Structured data refers to the kind of data that is organized and displayed in a database with

rows and columns, making it straightforward to work with. Examples of this include sales figures, names, phone numbers, and pretty much anything that can be categorized.

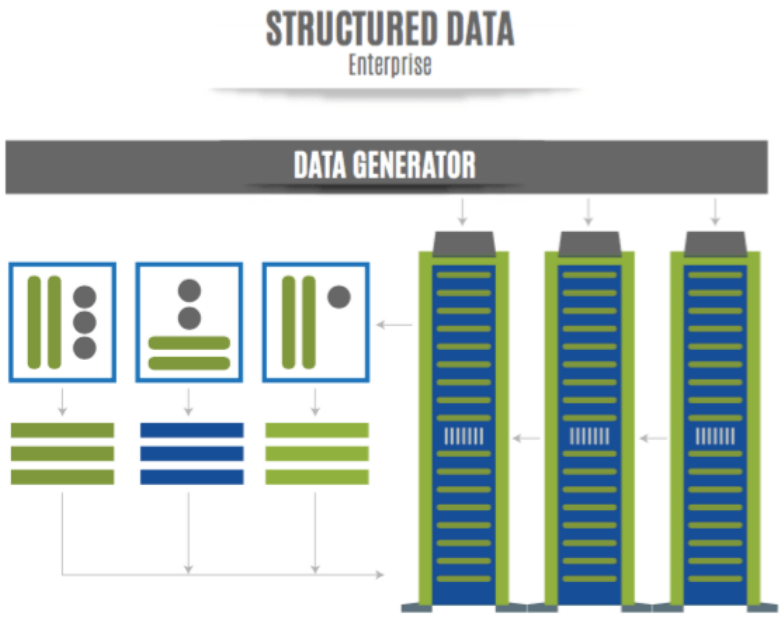


Photo Credit: The Executive's Guide to Big Data & Apache Hadoop

Unstructured data is the complete opposite. Due to its variability and unidentifiable internal structure, unstructured data cannot be analyzed by the conventional technologies.

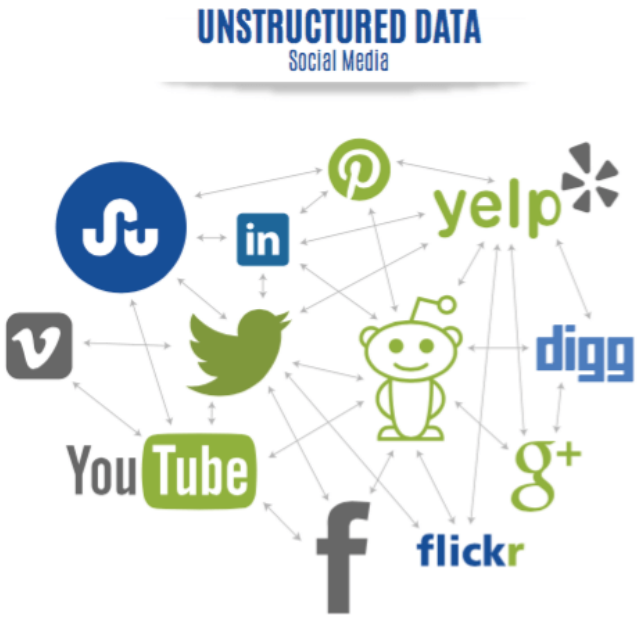


Photo Credit: The Executive's Guide to Big Data & Apache Hadoop

A few examples of 'unstructured data' are:

- *Social media posts*
- Images
- Emails
- Product reviews

When looking at social media posts, we see that most of the information can't be segmented into fixed categories due to the complexity and variability of the content. Social media users write about different subjects, in varying forms, making it hard to categorize them in a strict manner. Due to the increase in popularity of social media channels, new analytics tools and processes were developed to understand and extract value from this boom of unstructured data.

But why bother to extract this unstructured data?

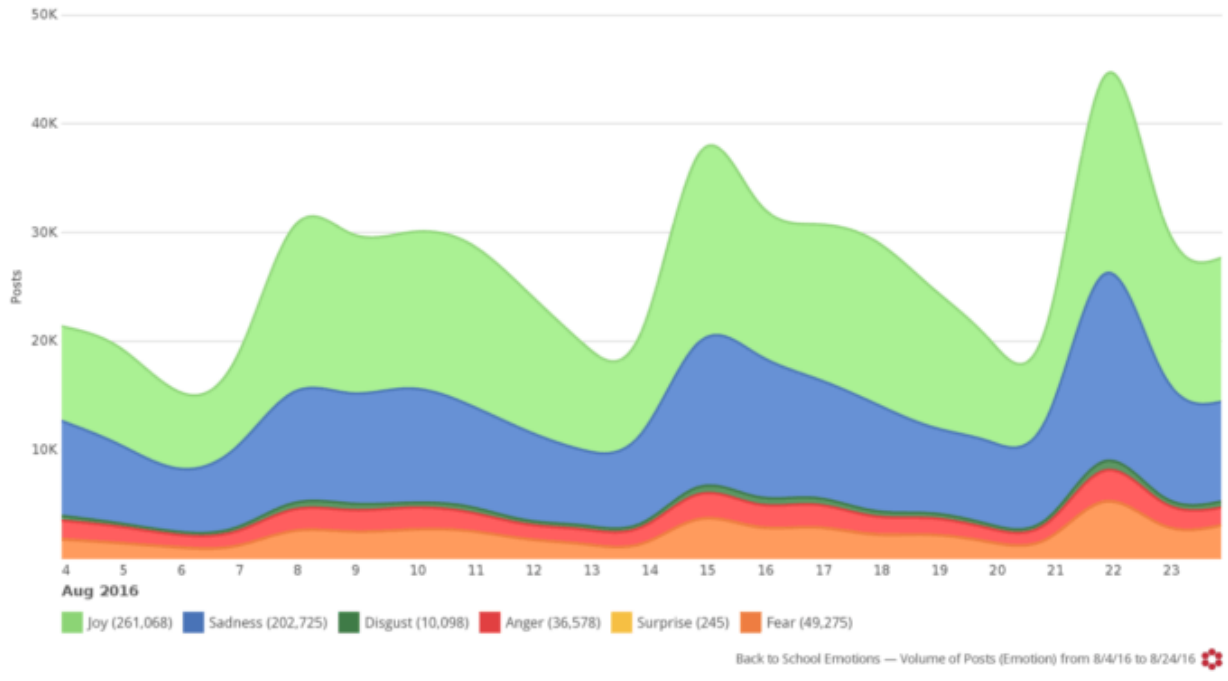
It can yield *deeper insights.*

Organizations in several industries are researching and investing in tools to extract meaning from this data and drive strategic business decisions, something hard to get from limited structured data. The value of unstructured data comes from the patterns and the meanings that can be derived from it; this includes identifying issues, market trends, or overall customer sentimenttowards a brand.

Two available solutions for the analysis of unstructured data are **Machine-Automated Natural Language Processing** and **Machine-Learning.**

Machine-Automated Natural Language Processing (NLP) and Machine-Learning

Natural Language Processing is a branch of artificial intelligence that allows a machine to understand the human 'natural' language. Therefore, the machine-automated solution tries to make sense of the data by processing statements and categorizing them in a systematic way. Applications of NLP on social data can identify general sentiment about a topic – either positive, neutral or negative – or it can even go as far as analyzing the universal emotions through emotion analysis.



Most analytics companies offer machine-automated features, but the problem is that the results can be inaccurate and not pertinent to the subject matter. Although this solution requires less setup time due to its automation, it risks providing irrelevant information to the user when analyzing conversations in different industries with particular dialects and slangs.

An example where a dialect or slang could be a problem is when a word such as *'wicked'* is used in different contexts. Meaning "evil or morally wrong", 'wicked' is widely identified as a word with a negative connotation, thus machine-automated processes would negatively categorize phrases containing this word. However, 'wicked' has a different meaning in Massachusetts; used as a positive word, it can mean "very" or "occasionally cool", making the outcome of the analysis very inaccurate and irrelevant to the search.

On the other hand, for more comprehensive text analytics, there's the machine-learning approach.

To better understand this concept, think of the online recommendations from Amazon; depending on your purchases, search history, your ratings, wish list, the interests of other similar customers, and more examples of what you're interested in, it will try to find items more relevant to your search.

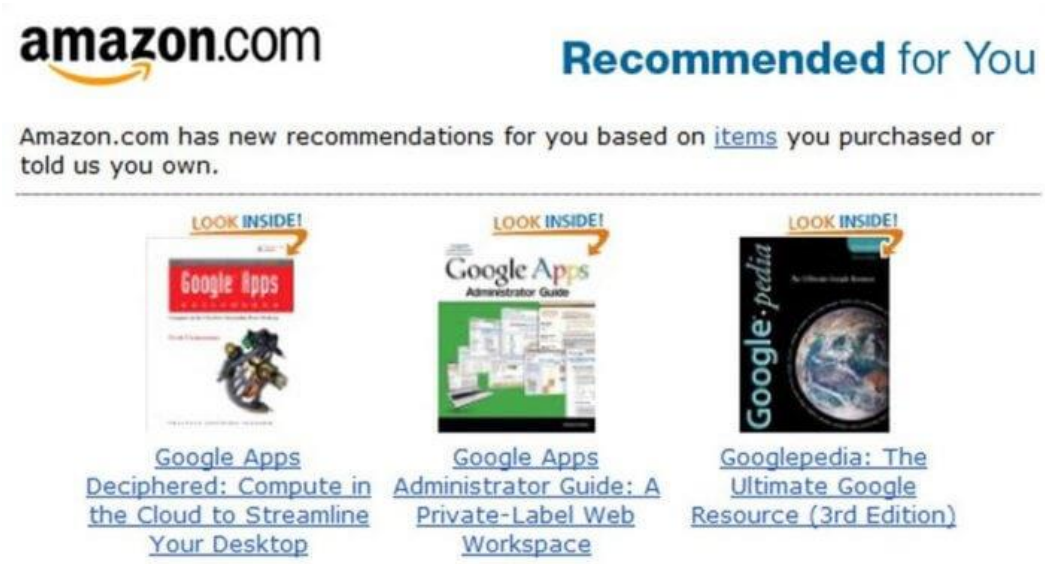
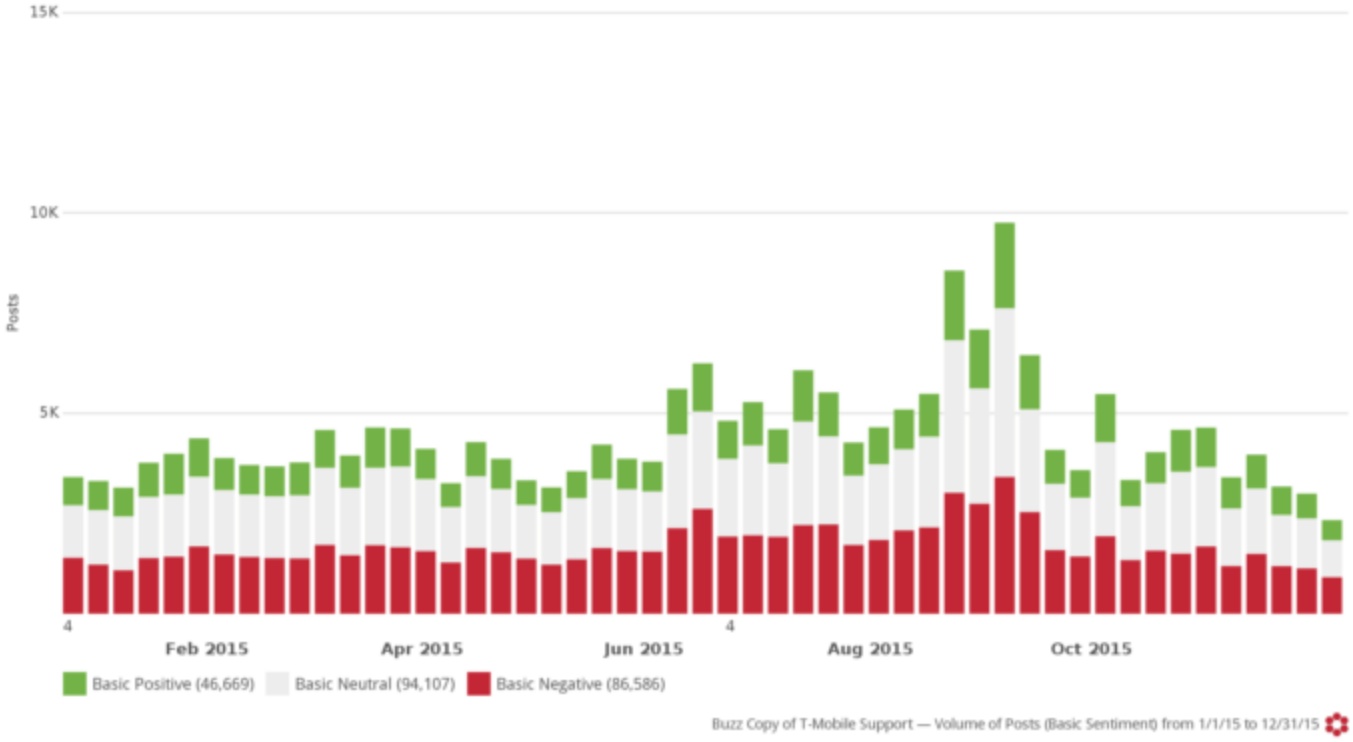


Photo Credit: Madmimi

Or even better, think of machine-learning as a *Gmail filter*. Fundamentally, filtering in Gmail is adding a label or tag to emails, so that it can count and group emails of the same kind together. If properly trained, Gmail's Inbox classifies emails into Topics like Social/Promotions/Updates etc. This model looks for patterns in the content of every email – such as keywords, phrases, authors – and assigns it to the most pertinent category; it doesn't follow pre-defined parameters.

Therefore, we can say that machine learning solutions allow tools to analyze multiple variables simultaneously, along with how they interconnect to form patterns. This option differs from machine-automated solutions *in many ways*; as the outcome will be more related to the question the user is trying to solve, machine-learning requires external knowledge and deep understanding of the conversation's subject matter to train the tools appropriately. Although training some posts to define each custom category may take some time, ***it will help the analysis tools to identify robust patterns and provide more relevant insights to the business question.***

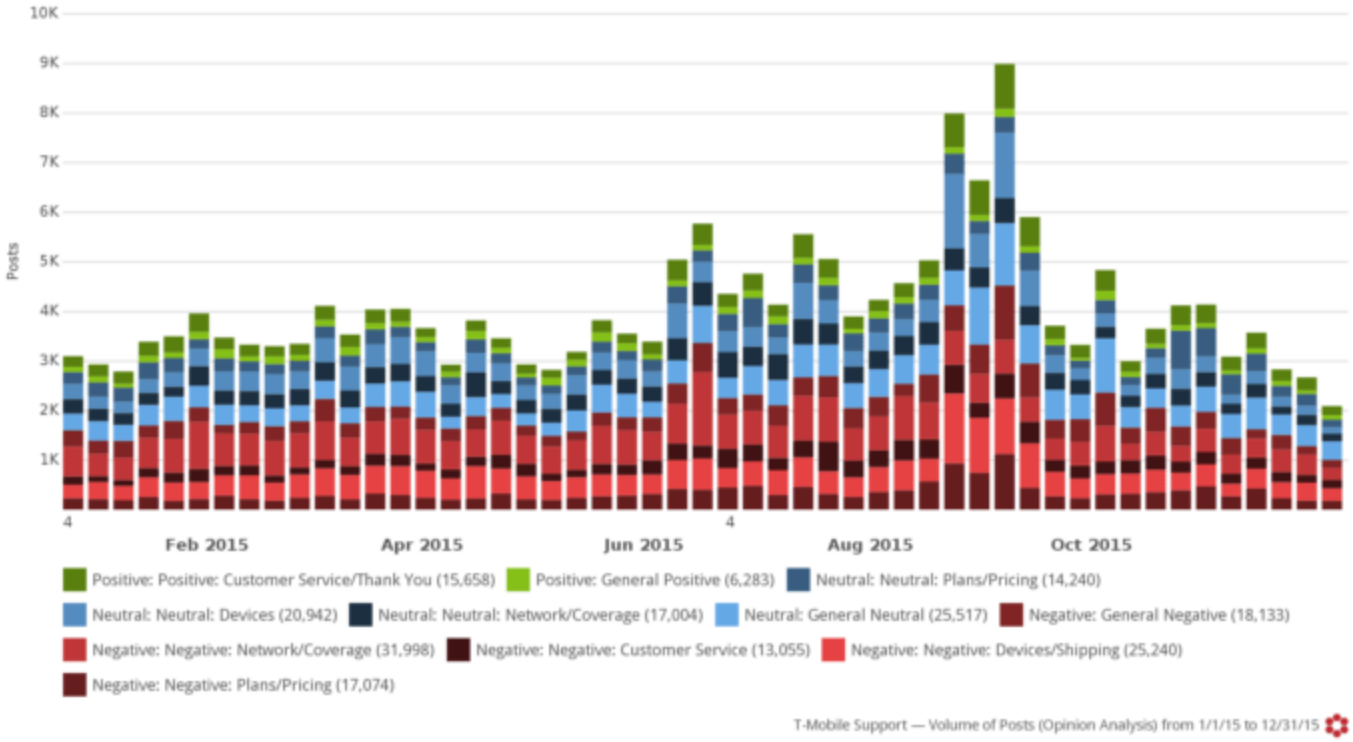
In the following example, we can see the analysis of T-Mobile's customer service on Twitter in 2015. Through machine-automated NLP, social posts can be categorized into general positive, neutral, and negative sentiment by identifying the keywords.



It's useful to know whether their customer service was overall positive or negative, but how can you extract more valuable, relevant insights from this data? How can you know why big part of the conversation is negative? This is where machine-learning comes in.

The following visual contains the same data and timeframe used to analyze T-Mobile's basic sentiment around its customer service. The main difference is that through machine-learning, you can create custom categories that will

explain the reason behind the positive, neutral and negative sentiment, providing actionable solutions that can applied.



Whether it is expensive plans, bad coverage, or insufficient customer service, organizations like T-Mobile can use machine-learning to understand the ‘why’ behind the negative sentiment towards their brand and make better-informed decisions in the future.

Custom categories in machine-learning processes allow the analysis to be more accurate and precise; some advantages include deeper nuance and meaning dependent on the user’s subject matter expertise and business context. With this solution, users can uncover insights to help them make more strategic decisions.