**BIG DATA**

Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques.

In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

**MEDIA AS A BIG DATA SOURCE**

Media is the most popular source of big data, as it provides valuable insights on consumer preferences and changing trends. Since it is self-broadcasted and crosses all physical and demographical barriers, it is the fastest way for businesses to get an in-depth overview of their target audience, draw patterns and conclusions, and enhance their decision-making. Media includes social media and interactive platforms, like Google, Facebook, Twitter, YouTube, Instagram, as well as generic media like images, videos, audios, and podcasts that provide quantitative and qualitative insights on every aspect of user interaction.

**CLOUD AS A BIG DATA SOURCE**

Today, companies have moved ahead of traditional data sources by shifting their data on the cloud. Cloud storage accommodates structured and unstructured data and provides business with real-time information and on-demand insights. The main attribute of cloud computing is its flexibility and scalability. As big data can be stored and sourced on public or private clouds, via networks and servers, cloud makes for an efficient and economical data source.

**THE WEB AS A BIG DATA SOURCE**

The public web constitutes big data that is widespread and easily accessible. Data on the Web or ‘Internet’ is commonly available to individuals and companies alike. Moreover, web services such as Wikipedia provide free and quick informational insights to everyone. The enormity of the Web ensures for its diverse usability and is especially beneficial to start-ups and SME’s, as they don’t have to wait to develop their own big data infrastructure and repositories before they can leverage big data.

**IOT AS A BIG DATA SOURCE**

Machine-generated content or data created from IoT constitute a valuable source of big data. This data is usually generated from the sensors that are connected to electronic devices. The sourcing capacity depends on the ability of the sensors to provide real-time accurate information. IoT is now gaining momentum and includes big data generated, not only from computers and smartphones, but also possibly from every device that can emit data. With IoT, data can now be sourced from medical devices, vehicular processes, video games, meters, cameras, household appliances, and the like.

**DATABASES AS A BIG DATA SOURCE**

Businesses today prefer to use an amalgamation of traditional and modern databases to acquire relevant big data. This integration paves the way for a hybrid data model and requires low investment and IT infrastructural costs. Furthermore, these databases are deployed for several business intelligence purposes as well. These databases can then provide for the extraction of insights that are used to drive business profits. Popular databases include a variety of data sources, such as MS Access, DB2, Oracle, SQL, and Amazon Simple, among others.

The process of extracting and analyzing data amongst extensive big data sources is a complex process and can be frustrating and time-consuming. These complications can be resolved if organizations encompass all the necessary considerations of big data, take into account relevant data sources, and deploy them in a manner which is well tuned to their organizational goals.

**SENTIMENT ANALYSIS**

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques.

Sentiment analysis allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback. For example, using sentiment analysis to automatically analyze 4,000+ reviews about your product could help you discover if customers are happy about your pricing plans and customer service.

**CHALLENGES WITH SOCIAL MEDIA DATA**

1. Grammar and Spellings: With users being too causal when posting on the web they tend to make a lot of mistakes in the semantics of the language and even the spellings of words. These are generally checked in the preprocessing stage of any application using these datasets.

2. Trustworthiness: The most important property of social data is the views of different users on different subjects, but there are many fake accounts being made to give fake views and reviews to either push or pull an entity on the platform.

3. Format: Every other social media site has its own style of posting data and also the way users post their data on these sites. Like people using # to tag subjects or using @ to refer to different users. Hence, it is important to study and understand each site differently.

4. Language: Social media sites provide options of using different languages to post views. There lie options to tackle this problem with either using translation mechanisms or building engines with respect to different languages.

**Unigram Model**

the assumption that the probability of a word only depends on the previous n words. This is known as an n-gram model or unigram model when n = 1. The unigram model is also known as the bag of words model

**Naive Bayes classification for feature reduction**

Due to its linear complexity, naive Bayes classification remains an attractive supervised learning method, especially in very large-scale settings. We propose a sparse version of naive Bayes, which can be used for feature selection. This leads to a combinatorial maximum-likelihood problem, for which we provide an exact solution in the case of binary data, or a bound in the multinomial case. We prove that our bound becomes tight as the marginal contribution of additional features decreases. Numerical experiments on text data show that the naive Bayes feature selection method is as statistically effective as state-of-the-art feature selection methods such as recursive feature elimination, l1-penalized logistic regression and LASSO, while being orders of magnitude faster.