CHAPTER 3

OPINION MINING AND PREDICTIVE ANALYTICS

3.1 Social Media Landscape

Social networking channels have unique strategies to engage the users and influencers for which there are no standard or traditional rules. Big data sources like Twitter and Facebook have unstructured data. These unstructured data has to be metamorphosed into a structured format for applying analytic techniques. Unstructured data is text heavy containing facts. They do not have a predefined data model that fits into a relational table. This is challenging as the traditional data mining techniques will not be able to dissect and get the required information. The volume of unstructured data captured and continuing to capture live is growing at an alarming rate. The semi-structured or unstructured data has to be converted into a structured form.

Social media posts millions of reviews every day. Capturing human emotion helps in the business understanding the sentiments behind the posts.

Sentiment analysis is the process of identifying the emotional tone from a series of words to understand opinions, feelings and attitudes expressed.

3.2 Scenarios

Traditional brand promotion activity to increase the brand awareness mainly depended on advertising channels through print or digital media or through direct mailing system. Social media opened a new avenue for cost-effective, responsive larger reach of audience. Social media helps in tracking customers and potential customers. The data on how many people saw the post and liked and reposted, give you a better picture of the reach of the activity. Unlike traditional marketing, social media marketing promotes information diffusion through

- Branding campaign to introduce business, promote a brand and connect to the target audience through an evolving network.
- Reputation management of the brand
- Meaningful connection to the audience with direct response
- Attract prospective customers to the website
- Measuring traffic to your site
- Identifying Metrics for measuring success.

Consumers share their opinions about any product as reviews, blog posts, and comments. The social network allows the users to broadcast any information to the public or closed group of friends or followers. Data-driven decisions are very much essential, for business to stay relevant in this

highly competitive environment.

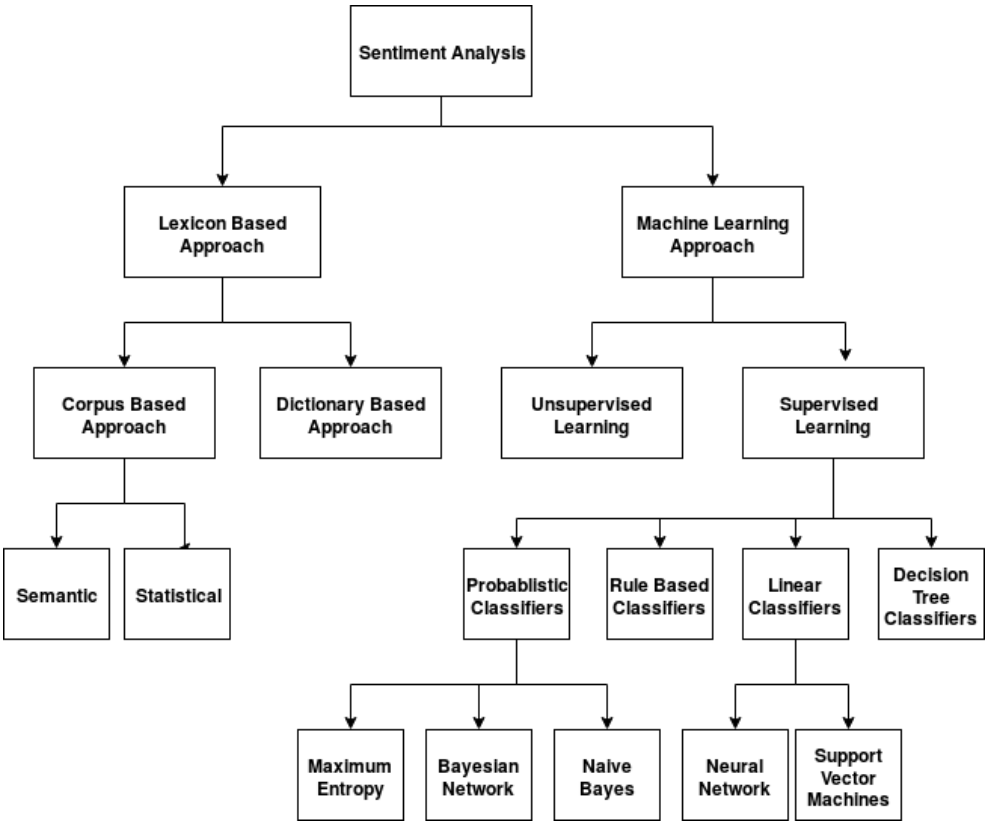
Understanding the visitor profile helps in promoting an individual's preferred brand with the content recommendation. This enhances the user experience, with improved and more personalised promotions. This helps in customer retention as well as capturing the new audience. Brand monitoring helps in cultivating good customer experience. Real-time conversations have to be constantly watched and monitored.

3.3 Sentiment analysis – Work Flow

The large temporal data generated through the conversations reflect the sentiment or reaction of people to an issue. Analysing the data deeply, valuable insights are discovered. Opinion mining, deals with the extraction of the opinion of people, from the large-scale data sets. Opinion mining looks into polarity and sentiment analysis is more of emotion recognition. Bay & Lee[120] analysed tweets to understand the trend of followers through Sentiment analysis. The investigation revealed the pattern of popular user's influence on audience. It was found that by measuring the sentiments of their audience, the positive or negative influence of the popular leaders can be measured. A Granger causality analysis found that there is a statistically significant causal correlation between the changes in Twitter sentiment and real-world landscape over the period.

Sentiment refers to the feeling or emotion, an attitude or opinion behind the conversation. The tone reveals whether the person is happy, sad, upset or

angry. There are many steps involved in sentiment analysis. Sentiment analysis helps in predictive analytics. The various techniques are pictorially represented in figure 3.1



Ref: http://doi.org/10.1016/j.asej.2014.04.011

Fig 3.1 Sentiment Classification Techniques

Sentiment analysis is used by business organisations to find how their brand or product is received in comparison to the competitors. Paying attention to the sentiment level gives opportunities in positioning your brand. There are

many steps involved in sentiment analysis. It can be also be used for predictions. The steps followed are given as a flowchart in figure 3.2

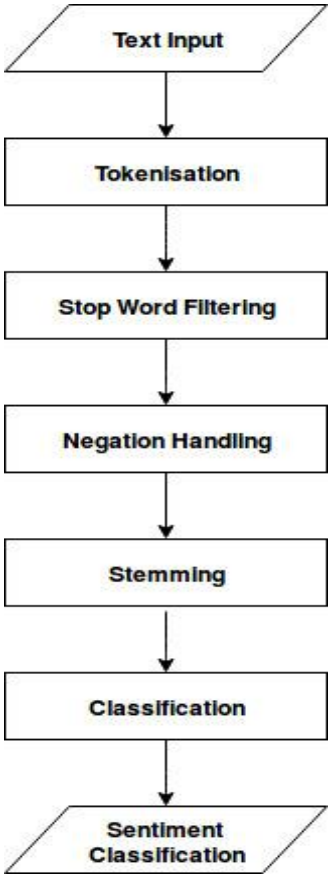


Fig 3.2 Sentiment Analysis - Workflow

The future research in Sentiment Analysis is discussed by Cambria et al.[121] pointing to the complexities involved. The current state of Sentiment Analysis is mainly dependent on text mining and by identifying positive and negative words. This text is processed and classified based on

emotions as positive or negative. Cambria et al.[121] stress on the fact the a deep understanding of semantic language rules is very much essential for 'feeling' emotions. The cognitive science and AI is developing at a pace that it can understand human emotions also. The availability of large volume of data along with cheap computational power and developments in NLP and related technology is an enabler to watch and listen to humans analysing their sentiments and emotions. This better understanding of emotion will help in creating more empathetic user experiences.

3.3.1 Data acquisition and Pre- Processing

The large temporal data generated through the conversations, reflect the sentiment or reaction of people to an issue. By analysing the data deeply, valuable insights are obtained. Opinion mining deals with the extraction of the opinion of the people from the large-scale data sets, captured from various sources like discussion forums, social media like Twitter, etc. Data capturing can be entirely automated or performed with the help of human input. A combination of automated methods and human processing will be the best option. People use social media platforms to express their opinion directly, and these social media conversations are constantly monitored. Data are from different sources and hence may be heterogeneous in nature including text, language, and domain. The data is pulled via authentication API. This need to be cleansed and unstructured data should be converted to the structured format.

We capture the Facebook posts and extract their content, in Data Capture

module with the help of Facebook‘s Graph API. The posts are tokenised to extract their keyword combinations. After that, feature selection is performed to keep only the n-grams that are important for the classification problem. The classifier is trained to identify the positive, negative and neutral posts. Pre-processing of the text – this is where the text is prepared for processing with the help of computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc.

After acquiring the data, the first step is cleansing. Data cleansing or data scrubbing involves the process of identifying and correcting (or removing) bad or corrupt and inaccurate records from a record set, table, or database. This is mainly used in databases to identify incorrect, or incomplete or irrelevant data. Data de-duplication is done to eliminate duplicate data. Language is checked. Outliers and noise are identified and removed.

Data cleansing include reformatting, de-duping, merging, and filtering. Data cleansing is basically the scrubbing and cleaning applied to raw data and getting it into a format that allows analysis.

3.3.2 Procedure for Analysing Sentiments with semantic search

Sentiment Sentiment analysis evaluates the written or spoken language and determine whether the expression is favorable, unfavorable, or neutral, and to what degree. By monitoring the online conversations, sentiment analysis along with text mining techniques helps to understand customer’s like and dislike about you and your brand. This is contextual. By regulaly

monitoring and reviewing the customer's feedback on your business, organisation can be proactive regarding the changing dynamics in the market place. Measuring the sentiment gives the overall feeling about a particular event or subject.

Advanced machine learning and Natural Language Processing techniques are used for performing Sentiment Analysis. Several text classifiers like Naive Bayes, Regression and Max Entropy are available to build a classifier capable of detecting Positive, Negative and Neutral sentiments in the posts.

Naïve Bayes classifier is simple to use because of the naïve assumption that each word is statistically independent of each other word. Sentiment analysis is subjective, whereas classification is objective. Subjective language can be more emotional with negation, sarcasm, idioms, and phrases. Bayes classifier is trained by counting the number of times the word appears in the class of documents. This can be extended to use bigrams, by tokenising the documents with paired words. Bigrams increase the entropy.

Information Extraction is the first step in the NLP process. Graph Aware NLP leverages Stanford CoreNLP [76] in order to perform the following text analysis operations:

1. Splitting of Sentences
2. Parsing
3. Tokenization

4. Lemmatization
5. Named Entity Recognition
6. Co-reference Resolution
7. Identifying Parts of Speech
8. Annotating

The process involves steps beyond tokenization; tokenizes, extracts lemmas and composes tokens that need to be together (i.e. dates, objects‘ names or organisation, geographic locations). This is a laborious and time consuming process.

The key steps are described in the following paragraph

Pre-select relevant words, from the NLP annotated text. Each document is tokenized and annotated using the GraphAware NLP Framework (these processed words are basic lexical units, also denoted as tags). A stop word list (configurable) and a syntactic filter are applied to refine the selection of the most relevant lexical units. The syntactic filter selects only nouns and adjectives, following the observation that even human annotators tend to use nouns, rather than verb phrases to summarize documents.

A graph of tag co-occurrences is created. Filtered tags are ordered based on their position in the document and co-occurrence relationships are built between adjacent tags, following the natural word flow in the text. This introduces the relations between syntactic elements of the document into the graph.

Natural language processing or linguistic algorithms assign sentiment values (positive, negative or neutral) converting the collected unstructured data into data sets. Text analysis is one of the most popular machine learning scenarios. Machine learning algorithms learn to capture useful information and reveal the relevant trends over time. Sentiment Analysis is an application of Classification problem categorised as Supervised Learning.

The NLP Framework has been extended to support unsupervised keyword extraction. It integrates the NLP functions available in different standard software packages like Stanford NLP and OpenNLP, and also data sources like ConceptNet5 and WordNet and recommendation engines. Recommendation engines are an alternative to search fields, helping users discover the products of interest to them. This matching is done by learning the properties of the items that a user likes; analysing what else the user may like; compare with likes and dislikes of other similar users, compute a similarity index between users and recommend items to them accordingly. GraphAware NLP can implement as a plugin for Neo4j and an external frontend for interacting with a Spark Cluster. It provides a set of tools, by means of procedures, background process, and APIs, which together provide a Domain Specific Language for Natural Language Processing.

Information Extraction (IE) is the processing of textual information for extracting main components and relationships. This can be entirely automated or performed with the help of human input. The best information

extraction solutions are a combination of automated methods and human processing.

The procedure of sentiment analysis with semantics is as follows:-

- Extracting sentiment
- Enriching basic data with ontologies and concepts
- Computing similarities between text elements in a corpus using base data and ontology information
- Enriching knowledge using external sources (Alchemy)
- Providing basic search capabilities
- Providing complex search capabilities leveraging enriched knowledge such as ontology, sentiment, and similarity
- Providing recommendations based on a combination of content/ontology-based recommendations, social tags, and collaborative filtering
- Unsupervised corpus clustering using Latent Dirichlet Allocation (LDA). To further lower the document's dimensionality. LDA is used because of its usage in text modelling which provides the interpretable lower dimensional representations of documents.
- Semi-supervised corpus clustering using Label Propagation
- Word2Vec computation and importing

Web pages are human readable. The vocabularies and semantics are created, to represent the information in the Resource Description Framework (RDF) in a directed graph format, for the machines to interpret. RDF is a data

model to store information. Web pages may embed RDF data, directly inside the HTML code.

Simple keyword matching is not enough to understand and predict. Contextual clues of words and phrases help to understand the hidden meaning. Search engines try to understand the relationship of data within the search landscape. It is very much essential to discover, and build knowledge bases (expressed as ontologies) more than just text mining for the understanding of the context. Most of the search engines use semantic web concepts for more accurate searching. Google‘s precise and fast searching algorithm Humming bird focuses on understanding the searches. They started initially utilising their own ontology, but later shifted to schema.org standards.

3.4 Ontology Modelling

Ontology is a collection of terms that identify the real things and relationship relevant to our domain. Ontology tools are built using graph database technology and W3C Linked Data standards. OWL class/subclass ontology is supported through user interfaces. Third party ontologies permit design, of complex knowledge organization schemas. Ontology modelling helps in creating a standard vocabulary, for describing resources and services which helps in semantic interoperability.

Massive data generated with high velocity and variety needs to be analysed for intelligence mining. It is impossible for human interventions considering

the volume, to examine and identify connections within data. Structure of raw data is defined, to make machines artificially intelligent. Ontologies define the abstraction model of the domain and identify the relationship among them. All concepts are defined. These defined models are machine understandable and shared. Ontologies have the following advantages:

- Consistency and standardization of data sharing model
- Knowledge discovery from various sources using intelligent support tools.
- Ontologies help in interdependencies. Concept-based opinion mining techniques build knowledge bases and analyse expressions to understand natural language opinions.

3.5 Computing Similarity

Once tags are extracted from all the news or other nodes containing some text, it is possible to compute similarities between them using content based similarity. During this process, each annotated text is described using the term frequency–inverse document frequency (TF-IDF) encoding format to evaluate the importance of a word in the document. Text documents can be TF-IDF encoded as vectors, in a multidimensional Euclidean space. The space dimensions correspond to the tags, previously extracted from the documents. The coordinates of a given document in each dimension (i.e., for each tag) are calculated as a product of two sub-measures: term frequency and inverse document frequency.

Term frequency describes how often a certain term appears in a document (assuming that important words appear more often). To take the document length into account, and to prevent longer documents from getting a higher relevance weight, some normalization of the document length should be done. The term frequency is divided by the document length (i.e., the total number of terms in the document) for normalization:

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document)

Inverse document frequency is the second measure that is combined with term frequency. It aims at reducing the weight of keywords that appear very often in all documents. The idea is that those frequent words are not very helpful, to discriminate among documents and more weight should, therefore, be given to words that appear in only a few documents. So we need to weigh down the frequent terms and scale up the rare ones, by the following computation:

IDF (t) = log_e (Total number of documents / Number of documents with term t in it).

Starting from such a vector representation, the similarity between vectors is computed using cosine. The procedure computes the similarity between all the annotated texts available and stores the first k similarities for each of them as relationships of type SIMILARITY_COSINE between all top k similar nodes

3.6 Providing Search Capabilities with Graph

Text processed during the annotation process is decomposed in all the main tags. Stop words, lemmatization, punctuation, pruning and other cleaning procedures are applied, to reduce the number of tags to the most significant. Furthermore, for each tag, its term frequency is stored to provide information about, how often a lemma appears in the document. Using such data and inspired by ElasticSearch scoring functions, GraphAware NLP exposes a search procedure that provides basic search capabilities leveraging tag information stored after text analysis. Search Tokens are the tokens obtained from the search query after text analysis. TF and IDF are computed for each tag, and other normalization factors are computed for each document in a query. GraphAware NLP extends basic search capabilities including ontologies, similarities, sentiment and so on. These features are combined with ElasticSearch queries to provide better custom results.

Based on data from DBpedia and WikiData, and smart machine learning algorithms, it recognises mentions of entities such as Person, Organisation, Location, key phrases, and relationships between them, as well as their relevance and confidence to the text. The site also provides a RESTful API to integrate the tagging service in your own system.

3.6.1 Algorithms behind the tagging service

The tagging service comprises of the following components:

- named entity disambiguation classifier - recognises the right "candidate" among overlapping annotations produced by the gazetteer and maps it to the right class and instance Uniform Resource Identifier (URI) in the knowledge base
- Named entity tagger - detects novel entities that are not present in the knowledge base and assigns them a generated class and instance URI
- Relation extraction rules - a set of rules which detects relationships between named entities
- Document classifier

Sentiment analysers can be built, to score the document, sentence or aspect using a combination of lexicon-based and machine learning algorithmic approaches. If sizeable training data is available, supervised learning is advisable, else unsupervised followed by a supervised classifier for typical text classification.

Build Model - The model is built on R, python or other analytical tools using the structured data transferred from the crawled unstructured data. NLP is used to tag each word with a sentiment and overall sentiment is calculated.

Train and Update - Train the model with new data set and update according to them for further accuracy.

Finding and classifying concepts – this is where mentions of people,

things, locations, events and other pre-specified types of concepts are detected and classified.

Connecting the concepts – The relationships between the extracted concepts is identified.

Sentiment Generation – NLP is used to tag each word with a sentiment, and overall sentiment is calculated as per the score.

Predictions- Sentiments of the review interprets in rating a product and helps in planning ways of improving.

Typically, for structured information to be extracted from unstructured texts, the following main subtasks are involved:

- Pre-processing of the text – this is where the text is prepared for processing with the help of computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc.
- Finding and classifying concepts – this is where mentions of people, things, locations, events and other pre-specified types of concepts are detected and classified.
- Relation Identification– relationships are identified between the extracted concepts.
- Standardisation – Extracted data is presented in a standard form.
- Noise Removal – this subtask involves eliminating duplicate data.

The identified keywords and key phrases are saved to the graph database via

a DESCRIBES relationship, between a keyword node, Keyword and the Annotated Text node. Label names are configurable.

The GraphAware TextRank procedure has many useful parameters allowing user-customization, but only one is mandatory: the Annotated Text node. No other argument needs to be specified - the TextRank performs very well out-of-the-box: The key phrases in the list appear very relevant to the NASA corpus. However, it is also important to note here that, out of those hundreds of key phrases, there are also clear misidentifications - phrases like overall historical or lower operation don‘t seem very helpful. One of the options that could help is, to improve the stop words list. However, there‘s also a more sophisticated approach discussed in the section below. The basic TextRank algorithm can be modified to leverage the typed dependency graph provided by Stanford NLP and integrated into the NLP Framework.

3.7 Predictive Analytics and Recommender System

Predictive analytics deals with extracting relevant information from data and further using it to predict direction and observance of performance patterns. It can be applied to any type of unknown data, whether it is in the past, present or future. It is passed through a data processing pipeline, which consists of multiple steps like Named Entity Extraction, Sentiment Analysis and a customized algorithm. Sentiment analysis uses text analysis and computational linguistics, to identify and extract subjective information in

source materials.

Machine learning algorithms can broadly be divided into supervised and unsupervised learning. In supervised machine learning, we train the algorithm with the labelled training set. This trained model is then being applied to new data, to predict a future condition. Supervised machine learning algorithms are used for predictive analytics. Statistical modelling techniques, like, time series forecasting and other forms of regressions are also used for predictive analytics. The difference is in the data generation process.

Predictive recommendation engines are the norm today. In recommendation engines, the best algorithms use a mix of content based recommendation, and collaborative filtering algorithms, to get a good result. This can be used for targeted marketing for specific groups of people. Here the profile of the customer is recorded, with which you can create a specific audience. Using clustering, similar groups of customers according to their purchasing patterns are matched with similarity index and reach the target segments.

……&&....