***Extracting Information from Social Network using NLP***

The information on the social network is unstructured and there is a need to extract the structured information for making use of the valuable information.

Natural language processing is employed to enhance the accuracy in visualizing the structured information that is speckled over the social network.

Analyses natural language text in order to extract information about different types of entities, relationships or events.

**INTRODUCTION**

Social Network is applicable almost in every field, to name a few; we have, voting mechanism for beauty pageant, political campaigns, product research & promotion via advertisements.

Data mining techniques are required for their capability of handling the three dominant properties with social network data namely; size, noise and dynamism.

Data mining techniques are designed to handle the voluminous data sets to mine significant patterns from data.

**NLP**

NLP techniques maps human language to machine language, it models the way user requests information to how computer or software understands it.

The encounter in Social Network monitoring is to extract and interpret ‘User Communication’.

NLP approaches that are essential for Social Network monitoring which are Automatic Summarization, Chunking, Part-of-speech tagging, Named Entity Recognition, Named Entity Disambiguation, Fact/Relation Extraction, Word- sense Disambiguation and Sentiment Analysis.

Automatic Summarization

Automatic Summarization is the process of reducing a text document with the help of a computer program in order to create a summary that retains the most significant points of the original document.

Two approaches to automatic summarization: Extraction and Abstraction.

*Extraction* refers to selecting a subset of existing words, phrases, or sentences in the original text to form the summary.

*Abstraction* builds an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate.

Three basic steps of automatic summarization, Analysis, Transformation and Realization

Chunking

Basic technique used for entity detection.

Selects a subset of the tokens rather than tokenization that omits whitespaces.

Removing a sequence of tokens from a chunk is refereed as Chinking.

Parts-of-speech tagging

Piece of software that reads text in some language and assigns parts of speech to each word such as noun, verb, adjective to name a few.

Dictionaries have categories of a particular word which implies that a word may belong to more than one category. Taggers employ ‘Probabilistic Information’ to solve this ambiguity.

Named Entity Recognition

A subtask of information extraction that seeks to locate and classify Named Entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Hand-crafted grammar-based systems typically obtain better precision.

Currently statistical models are preferred as this approach initially uses training data against the model, followed by preparation of statistics

Named Entity Disambiguation

The task of linking the identity of entities available in the text

Distinctive from named entity extraction as it identifies not the occurrence of names but their reference.

Fact/Relation Extraction

Once named entities have been identified in a text, we can then extract the relations or facts that exist between specified types of named entity.

Objective is to detect and distinguish the semantic relations between entities in text or relations and fill it in a predefined template using the entities.

Word Sense Disambiguation

An open NLP and ontology subject that identifies the correct sense of the word in a sentence where multiple meanings of the word exist.

Identification of the aspect of a word is difficult for a machine to understand. This methodology provides a mechanism to diminish the ambiguities of words in the text.

Sentiment Analysis

An NLP process which identifies, extracts, enumerates the attitude of the user to the information that is provided by the user in a free form text.

Extensively used in processing survey form, online reviews and social media monitoring.

Returns the identified sentiment with a numeric score from 1.0 to -1.0 where 1.0 means strongly positive and -1.0 means strongly negative

**OPEN SOURCE NLP LIBRARIES**

Apache OpenNLP, Natural Language Toolkit (NLTK), Standford NLP, MALLET

**CHALLENGES IN NLP**

1. *Informal language:* Users post texts in an informal language which is noisy

2. Part-Of-Speech tags make the Information Extraction from social network more challenging.

3. *Short contexts:* minimum length prompts the users to use more abbreviations to precise more information in their posts.

4. *Noisy sparse contents:* The users’ post on social network does not always contain useful information. To purify the input posts stream, Filtering is required as a pre-processing step

5. *Information about entities:* entities are not always contained in the knowledge Base. There is a need of new Suit of Information Extraction from Social Network posts.

6. *Uncertain contents:* Not all information is trustworthy on the social network. The uncertainty involved in the extracted relations/facts is difficult to handle.

**TEXT MINING**

Refers to the employment of data mining techniques which automatically discover and extract information from unstructured text documents and services.

Searching with the help of text mining is a way of retrieving and searching on a social searching engine that mainly searches user-generated content such as news, videos and images related search queries on social media like Facebook, Twitter etc.

The text mining approach consists of four steps, which include Data Collection, Pre-processing, Generalization and Analysis

Data Collection

Process of gathering and measuring information in a systematic manner, which then enables one to answer relevant questions and evaluate outcomes.

There are huge numbers of users who access historical data at a particular time and it becomes difficult and expensive for social network to gather large amount of data. So, Summarization maintains all important data and further discards the insignificant data.

Pre-processing

Processing of raw data to deliver a podium for data analysis

Purpose of it is to classify raw sentences into sentences which can be read by the machine.

The text is cleaned and delimiters are removed with the help of some pre known list of stop words which are not useful to classify the meaning of the sentence. The text and its characteristics are pointed in an attribute value table.

Users enter the social text in a free form and therefore it is a challenging task to classify that data. Just to be sorted out from this challenge, part-of -speech tagging and Named Entity Recognition are used.

Generalization

Involves the multiple patterns at the text of the preprocessed texts.

Deals with developing algorithms to ascertain stimulating, unforeseen and unusual information form the patterns in the text document.

Association, correlation, classification, cluster analysis form the strong foundation of data mining chores.

Analysis

Deals with the validation and interpretation of the generalized data pattern.

Degree identifies the “connections” between the users.

Centrality determines the behaviour of individual user in the associated network.

Indegree claims the individual user as the central identity; centrality of other users is based on their relation to the user

Outdegree the interaction of the user with others is the main focus point.

Sociogram is the representation with the limited boundaries of the connection in the network is the point of analysis.

**APPLICATIONS OF TEXT MINING IN SOCIAL NETWORK**

1) Keyword Search

A set of keywords are used to identify the social network nodes which are close to the query result. Content and Linkage behavior plays an important role in order to determine the query output. It provides an effective method for accessing structure data.

Query Semantics, Ranking Strategy and Query Efficiency are the major concerns to perform keyword search in social networks.

2) Classification

The nodes in the social network are associated with labels which are used for classifying the network.

The major issue in classifying the labels of the social network is the non-standard vocabulary and the noisy information associated with the labels as the labels of the social network are often sparse.

3) Clustering

Area where set of nodes are used to determine the similar content for the evolution of clusters.

K-means is the widely adopted technique where initial value of k is specified and clusters are built iteratively around that value.

Linkage of clusters is an important factor and when combined with content can classify the social network which results in better clusters.

Heterogeneity of the social network is the major concern that makes the algorithmic design more difficult to provide a viable solution.

4) Linkage based Cross domain learning

The linkage information between multiple domains of social networks provides transfer of knowledge across various kinds of links.

Major concern is the amount of training data available from multiple social networks. An effective learning process can be leveraged from the various domains of the social network.

**CONCLUSION**

Traditional text mining techniques are not popularly used in social network monitoring. The combination of text mining and web mining techniques should be incorporated to analyse a social network monitoring system. NLP Techniques will enhance a user friendly search by the Social Network user.