A
Major Project-II Report
On

CONCEPT BASED TEXT CLASSIFICATION

Submitted in Partial Fulfilment of the Requirement for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

ARUNIMA JOSHI
2K14/CSE/05
Under the Esteemed guidance of
DR. AKSHI KUMAR

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road,
Delhi-110042.
JUNE, 2016

CERTIFICATE

This is to certify that the work contained in this dissertation entitled **"Concept based Text Classification"** submitted in the partial fulfilment, for the award for the degree of M.Tech in Computer Science and Technology at **DELHI TECHNOLOGICAL UNIVERSITY** by **ARUNIMA JOSHI, Roll No. 2K14/CSE/05**, is carried out by her under my supervision. This matter embodied in this project work has not been submitted earlier for the award of any degree or diploma in any university/institution to the best of our knowledge and belief.

(Dr. AKSHI KUMAR)

Project Guide

Assistant Professor

Department of Computer Engineering

Delhi Technological University

DECLARATION

I hereby declare that the major Project-I1 work entitled "**Concept Based Text Classification**" which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of degree of Master of Technology (Computer Science and Engineering) is a bonafide report of Major Project-1I carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Arunima Joshi

2K14/CSE/05

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Dr. Akshi Kumar for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to her for the support, advice and encouragement she provided without which the project could not have been a success.

Secondly, I am grateful to Dr. O.P.Verma, HOD, Computer Science & Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out. Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Arunima Joshi
University Roll no: 2K14/CSE/05
M.Tech (Computer Science & Engineering)
Department of Computer Science and Engineering
Delhi Technological University
Delhi – 110042

ABSTRACT

The shift from Web 2.0 to Web 3.0 has significantly changed the perception of users for the internet and the Web. Web 2.0 has improved information sharing among the users, the contribution and collaboration of users, and Web 3.0 has improved the structure and representation of data. Web 3.0 (Semantic Web) is all about the concepts which relates more to real-world entities, which proves to be more realistic and practical.

One of the most popular applications of Web 2.0 is blogging and its services. For example, Twitter has evolved as a great platform to share opinions and views on anything and everything related to daily life. As a result, these blogging websites have emerged as rich database for sentiment analysis and opinion mining. However, due to the nature of tweets (syntactically inconsistent), text-based sentiment classifiers fail to prove efficient. Concept based text classifiers are now-a-days gaining popularity and proving to be more efficient in such situations.

Sentiment Analysis is inherently a prominent Text Classification problem. In order to improve efficiency of text classifiers, concept based techniques can be used. So, the inclusion of conceptual features of Semantic Web into the Text Classification problem can benefit the process of analysing the opinions of users. Therefore, in this research work, ontology-based techniques are used as concept based text classifier, to classify the text (tweets) more efficiently. In this approach, ontology is used to extract features or attributes on which tweets are to be analysed and accordingly score is given. The domain chosen is Ministries of Indian Government.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The shift from Web 2.0 to Web 3.0 has significantly changed the perception of users for the internet and the Web. Web 2.0 has improved information sharing among the users, the contribution and collaboration of users, and Web 3.0 will focus on improvement of the structure and representation of data. Web 3.0 (Semantic Web) is all about the concepts which relates more to real-world entities, which proves to be more realistic and practical. Whereas in Web 1.0, users were not active in terms of discussions. They just could view the information, but could not actively reply back or contribute to it. Of all the applications of Web 2.0 such as social networking sites like Facebook, wikis, blogs, multi-media sharing sites like YouTube, and other applications, blogging has gained popularity gradually.

Blogging initially received comparatively less attention, but now it has become very popular communication platform. Within this, a term which has also gained popularity is Micro-blogging. It let users to post their views, queries, experiences or opinions on any topic with limited content size. For example, Twitter has evolved as a great platform to share opinions and views on anything and everything related to daily life with each tweet size up to 140 characters. This character limit is very helpful because users can post tweets which are concise and more expressive than any other content. Therefore, tweets become rich content for mining the sentiments of people about any topic.

Opinion mining, also known as Sentiment Analysis, is the process of analysing the sentiments of people about a topic. It focusses on determining the viewpoint of a user related to a topic or a document's polarity as a whole. The viewpoint can be due to his/her evaluation or analysis or any judgement, present state of the user like his/her emotional state at the time of writing, or any false intentions like to defame someone which may have negative impact [1]. Sentiment Analysis is inherently a prominent Text Classification problem. In order to improve efficiency of text classifiers, concept based techniques can be used. So, the inclusion of conceptual features of Semantic Web into the Text Classification problem can benefit the process of analysing the opinions of users. As the domain is Ministries of Government of India, sentiment analysis would be useful for the government to monitor the schemes they have introduced, to

get an overall view of how citizens think about their work or about any other topic related to the country which can be used to improvise the growth of country [2].

1.1 MOTIVATION

Motivation for this research work came from the increasing popularity of Web 3.0 and the "Digital India" initiative, recently proposed and projected by the Government of India. This initiative requires continuous monitoring and tracking of the scheme and how this scheme is being received by people, like their opinions and viewpoints. As twitter has become the most popular microblogging site where billions of people post their opinions about a topic, in order to monitor the performance, one efficient way is to do sentiment analysis on twitter posts so as to get an overview or in-depth analysis depending upon the method used for sentiment analysis. As Web 3.0 is evolving rapidly and as it supports a well-structured format for data representation, it would be efficient to use data in form of concepts for the techniques used in text classification so as to have a more detailed and systemized analysis of the topic. This research work basically tries to inculcate semantic web features with sentiment analysis techniques. Hence the idea is to envision "digital governance" by virtue of social web adoption, where government can take advantage of social platforms involving huge user participation.

1.2 SCOPE

This project could be used by Government of India to track their schemes, policies, practices, rules and to monitor their performance. It would help the Government as there does not exist such work or model in this domain till date. Also, the domain can be changed to any others and the same whole sentiment analysis can be performed in that. Also if there are some other techniques for classification which can optimize the system, then they can also be used as there are different modules in this work which can be changed with some other module which can make the system better.

1.3 RESEARCH OBJECTIVES

Research Question:

"Can Semantic Web concepts empower the process of sentiment classification and analysis on microblogging sites?"

In order to classify the text (the opinions/ viewpoints of users) taken from the microblogging sites (specifically twitter) based on some concepts, the above question can be divides into following questions and all of these would be answered by this research:

- How the Semantic Web characteristics are useful for text classification?
- How various features can be extracted automatically for sentiment analysis using Semantic Web framework?
- How much difficulty level is there to converge Web 2.0 applications with Web 3.0 framework?
- Finally, what are the applications where this model can be used?

Therefore, the research objectives are:

1. To seek how Semantic Web features, Web 2.0 applications and sentiment analysis come up together.
2. To propose an approach which involves feature extraction using concepts and bag of words based method to find the polarity of the tweet.
3. To find out how this approach will serve for real-life applications.

The objective of this thesis is to perform text classification based on some concepts as features to determine the sentiment of the users about various topics, together as a whole or separately as an individual, depending upon the need.

1.4 ORGANIZATION OF THESIS

This thesis is structured into 5 chapters followed by references.

Chapter 1, as discussed, presents the research problem, objectives, justifies the need and use of using the approach and the research questions.

Chapter 2 presents the pre-requisite knowledge or background for this thesis and provides the novelty of our work.

Chapter 3 provides the details of the proposed framework, the methodology employed, the platforms used and outlines the model of Sentiment Analysis done on Ministries of India.

Chapter 4 presents the results obtained after performing sentiment analysis on tweets with some sample tweets as illustration.

Chapter 5 presents the conclusion of this research work and the future work that can be done on this system.

CHAPTER 2: LITERATURE SURVEY

2.1 EVOLUTION OF WEB

The World Wide Web was invented by Tim Berners-Lee in 1989. While working at CERN, Switzerland, he created the first web browser in 1990. The World Wide Web is a global knowledge sharing platform where users can access data via computers connected to the Internet. People generally take Web and Internet as synonyms but these two terms are totally different. Internet provides the service to users to stay connected to Web and read and write the data.

2.1.1 WEB 1.0

Web 1.0 is called as the first instance of the World Wide Web, which consisted of Web pages and were connected by hyperlinks. Although there is no exact definition of Web 1.0, it is more of a static Web, basically consisting of static websites which did not provide any interactive interface or content. According to Berners Lee, Web 1.0 could be called as "read-only web." In other words, users could search for and read the information using this web. Very less user interaction of users was there. Initially the goal was to make the information available to users. For example, any website owners wanted to establish an online presence by making information about them available to anyone at any time.

2.1.2 WEB 2.0

Web 2.0 is the second generation of the World Wide Web. It allows users to contribute and share information online. Web 2.0 is the evolution from static Web to a more dynamic Web which is more structured and interactive.

Web 2.0 has another improved feature of open communication and open sharing of information. The components of Web 2.0 includes blogs, wikis, social networking sites and Web services. Web 2.0 is the current version of online technology with features as greater user

interactivity and collaboration, more network connectivity and enhanced communication channels.

Another important difference between Web 1.0 and Web 2.0 is due to the social nature of Web 2.0. It enables interaction, content – sharing and collaboration. Various social media sites and applications of Web 2.0 include wikis, microblogging, forums, bookmarking. Web 2.0 is more of a "Read-Write Web". Any user can read the content and at the same time can write something on the website.

2.1.3 WEB 3.0

Third generation Web, Web 3.0 is still in research and would be akin to "Read-Write-Execute Web". Web 3.0 will solve the problem of vague Web versioning of Web 2.0. Web 3.0 will support machine-to-machine interaction over the Internet. Web 3.0 guarantees the ability of applications that can interact directly by relating semantics and web services.

One important thing to note is that no upgradation is needed or downloading new software or anything like that. Web 3.0 is just an abstract idea or the next fundamental change as in how to create the websites according to Web 3.0 semantics and, more importantly, how users can co-operate with these websites. It's a rumour that it's impossible to implement Web 3.0. In order to change Web first generation to Web 2.0, researchers devoted about their 10 years of effort to this and it may take this much long or more for the next fundamental change to occur.

Some features of Web 3.0:

- *The Artificially Intelligent Web*: Use of artificial intelligence in many applications is thought to be next big revolution on the web. Social Media has one benefit that it factors in human intelligence. For example, if social bookmarking is used as search engine, it can give more intelligent results instead of using Google. But the results can be influenced due to the human factor. Some users could give false comments or could vote for a particular content in order to make it more popular. So, if artificial

intelligence can be used to differentiate between true and false, it could give better results with elimination of false elements.

- *The Semantic Web*: A lot of research is going on related to idea of Semantic Web, a web where information is well structured, categorized in such a way that a machine and a human, both are able to understand it. Web 3.0 is considered as a combination of semantic web and artificial intelligence. The structure of information on semantic web will tell the machine about the data and artificial intelligence can use the data for many applications.

- *The World Wide Virtual Web*: This is an abstract idea, but belief is that it might lead to a web based on a virtual world. For this to happen, it is required that all the websites have some particular standards and framework that would allow machines to understand it and use the information efficiently.

The structure of web is such that visiting a page is easy but understanding it is impossible by machines. The search engines don't understand the keywords and their context with the page on which they give results while searched. The difference Web 3.0 will make is that machines will be able to search, look for and interpret the information on the pages using some agents. These agents are software programs which crawl through the web for the information and will give the relevant results. This interpretation of data by machines is possible due to the collection of data using Ontologies. Ontology is a hierarchical structure which defines various entities and their attributes, and their relationships between them.

In order to make Semantic Web work effectively and efficiently, ontologies need to be developed in such a way that they are detailed and complete in all manner in particular format that machines will understand. It consumes a lot of time and effort to develop an ontology because first it needs to be conceptualised which takes more percentage of time.

Figure 1

2.2 SENTIMENT ANALYSIS

Sentiment analysis (also known as opinion mining or text classification) refers to the process of identifying and extracting the sentiments / opinions of the document or text. It can be widely used for variety of applications in the field reviews, blogs, social networking and the domain ranging from marketing, business to customer service. In general, the purpose of sentiment analysis is to analyse the opinions or views of people related to some subject or to attitude of a speaker or a writer with respect to some topic or of a whole document. The opinion of a person can be his/ her belief or judgement or it can be due to mental state or the bad intentions [3].

2.2.1 TYPES OF SENTIMENT ANALYSIS

The basic job in opinion mining is categorising the polarity of the data at the feature, sentence or document level as whether the opinions expressed by the user are positive, negative or neutral. Beyond these three classes, further classification looks into more deep emotional states like “happy”, “sad”, “angry”, “depressed” and more.

Pang [5] and Turney [6] employed different methods to detect the overall sentiment of movie or product reviews at document level respectively as initial stage of work. Pang [7] and Snyder [8] attempted in-depth analysis beyond this polarity: Pang and Lee [7] expanded this basic classification of movie review into just three classes, to rating scales i.e. 3 or 4 star scale, while Snyder [8] predicted ratings for various aspects of the given restaurant from given restaurant reviews. Although in the most classification methods, due to the nature of binary classifier, the neutral class is overlooked, thee classes must be discovered for every polarity problem, as

suggested by several researchers. In addition to this, there are some classifiers like Max Entropy [9] and SVMs [10] which are benefitted from the use of neutral class.

To determine the sentiment, a different approach could be the use of scaling system in which words having sentiment as negative, positive or neutral can be given a rating on a scale from -10 (most negative) to +10 (most positive). It is useful in terms of relating the sentiment of the term to its environment like level of the word in the sentence or a document. Each concept in the specified domain is associated with a score which is based on the approach the sentiment words correlate to the concepts and their scores [11]. This will enhance the interpretation of the sentiment given to the concepts, relative to the domain.

2.2.2 SUBJECTIVITY/OBJECTIVITY IDENTIFICATION

It is the process of categorizing the given text into one of the two classes: subjective or objective [12]. At times, this task could be difficult than the polarity identification [13]. The context determines the subjectivity of the words or document and there could be some subjective content in an objective document. Su [14] mentioned that the results largely depend on the how the subjectivity used is defined while annotating document. On the other side, Pang [15] indicated that if objective sentences are removed from the document before classification, then the outcome in terms of polarity was better and improved performance.

2.2.3 FEATURE BASED SENTIMENT ANALYSIS

Feature based sentiment analysis refers to the process of determination of the sentiments or the opinions stated on diverse range of features of objects like phones, websites, restaurant etc. Feature could be an attribute or property of that object like the camera of phone, the food of restaurant or the service of website. This type of sentiment analysis has an advantage that it is able to depict only those characteristics of the object which are of interest. It is also possible that different features may give different sentiments e.g. a restaurant may have good food but bad atmosphere. So if classification is done on the basis of food as feature then it will classify that restaurant as good hotel but if the same is done on the basis of atmosphere then it will classify the hotel as bad hotel. So this problem includes various sub problems, like

identification of the relevant objects or entities, then obtaining the features or attributes, and then determining the sentiment of the opinion related to the features, as positive, negative or neutral. There could be a way of automatically identifying the features using some syntactic models.

2.2.4 METHODS AND FEATURES

There exists some methods of sentiment analysis which can be broadly categorized into three classes: statistical methods, knowledge based methods and hybrid techniques [4]. Statistical methods take into account the machine learning techniques such as SVMs, semantic analysis, semantic orientation, bag of words etc. Knowledge based methods performs the process of classification of text on the basis of presence of explicit words which have affect state like happy, angry, worried, sad, bored etc. There are other methods which aim to identify the bearer of the sentiment (the person who writes the opinion) and the target (the person or entity for which the sentiment is written). Hybrid approaches works on the amalgamation of knowledge based methods like semantic networks or ontologies and machine learning techniques to discover the meaning or sentiment expressed in refined structure i.e.by analysing the concepts which are not explicitly defined but they link to other concepts implicitly [4].

There are some open-source tools that use machine learning, some statistics and natural language processing methods which will automate the process of opinion mining on huge set of data like reviews, news available online, web blogs, social networking sites, group discussions etc. Also, knowledge based methods use resources which are available publicly such as SenticNet, SentiWordNet, OpenDover etc. to get the semantic and other associated information. There are some approaches using which visual content can also be analysed. The first approach for this purpose is SentiBank which takes into account the representation of visual content in terms of noun-adjective pair.

2.2.5 EVALUATION

The system can be evaluated by the accuracy of output as in how well it is oriented with the human judgements. The two measures for accuracy evaluation is Precision and Recall.

Precision is the percentage of retrieved results which are relevant. Recall is the percentage of relevant results which are retrieved. According to research, the people who gives rating, agree 80% of the time. So, the system with 70% accuracy is working well and nearly as human beings. System can be evaluated using more measures, but it is a complex issue to evaluate system performing sentiment analysis. If the output of sentiment analysis is on a scale than just 2-3 classes, the better measure metric would be correlation which checks for the closeness of the predicted result to the expected result.

2.3 SENTIMENT ANALYSIS AND WEB 2.0

Importance of Sentiment analysis is increasing day by day due to the growth of social media like social networking sites, blogs, discussion forums etc. Opinion given online in terms of ratings, reviews and recommendations etc., has become a virtual world is used for various purposes like marketing the products, looking for new prospects and managing the status. Many people are going for sentiment analysis to grow their businesses in order to systemize the process of filtering out the unwanted objects, understanding the text, recognizing the relevant information and using it efficiently. While Web 2.0 allowed to contribute or publish the data anytime anywhere, the next phase Web 3.0 may be all about web mining of all the data. There are many research groups which are working on understanding the subtleties of the sentiments on the web using sentiment analysis.

There is a hitch in this process that the algorithms used in sentiment analysis makes use of simple and basic terms for expressing the opinion. But, due to diverse cultural languages, unlike contexts and other factors, it is very difficult to classify the text into pro or con state. Generally, human beings disagree on their own opinions about a topic that it is hard to make machines give the right opinion. It becomes harder to do sentiment analysis when the text is really short. Although the text might be short and it might create a trouble, opinion mining done on microblogging sites has proved that Twitter is one such site which can be used as valid online gauge for opinion mining [4].

2.4 SEMANTIC WEB AND ONTOLOGY

World Wide Web Consortium (W3C) extended the definition of Web to Semantic Web through its standards and framework by promoting same data formats and protocols throughout the Web. According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". Semantic Web uses languages to publish the data: Web Ontology Language (OWL), Resource Description Framework (RDF), and Extensible Markup Language (XML). HTML language is used to define pages and hyperlinks between them whereas these 3 languages can define any type of thing such as people, their characteristics, any place or a thing. These languages are used to give descriptions that may enhance the web contents. So, the content will be like descriptive data in Web database. These descriptions readable by machines, would allow content managers to associate semantics to content that is to add the meaning to the data or to define the format of the data according to the users.

2.4.1 CHALLENGES

Semantic Web has many challenges which are vagueness, vastness, uncertainty, inconsistency, and deceit.

- *Vagueness:* Vagueness refers to terms with no clear meaning. For example, words like "young" can't tell how much young. It may arise due to the nature of queries of users or from concepts structured by content managers. Vagueness can be dealt with fuzzy logic technique.
- *Vastness*: There are billions of pages available on Web. Huge data can be dealt with any automated system.
- *Uncertainty*: Uncertainty is related to a concept having uncertain description or value. To deal with uncertainty, probabilistic reasoning technique is used.
- *Inconsistency*: During development of large ontologies or when ontologies are merged, there are some logical contradictions that may arise which are inconsistencies. Inconsistency can be dealt with defeasible reasoning and paraconsistent reasoning.

- *Deceit*: Deceit is that situation where any information is created such that it intentionally misleads who uses that information. It can be alleviated by cryptography techniques.

2.4.2 COMPONENTS OF SEMANTIC WEB

Semantic Web is generally used to denote formats and related technologies that are used to enable it. These technologies enable the organization and recovery of collection of linked data in a structured format and this will further give a formal representation and annotation of some domain-specific concepts, their attributes (features) and their inter-relationships. All of these are defines in W3C standards and it contains following components:

- Resource Description Framework (RDF)
- Simple Knowledge Organization System (SKOS)
- RDF Schema (RDFS)
- SPARQL, query language
- Web Ontology Language (OWL)
- Rule Interchange Format (RIF)

Resource Description Framework is a structure which describes the data models, some objects and their relationships. It is the basis of Semantic Web. RDFS broadens the definition of Resource Description Framework. It is a dictionary for defining the schema, the properties and the classes and objects of RDF resources and also includes semantics for generalizing relationships of these classes, objects and properties. RDFS is more of a generalized language, but OWL is more specific in nature. It adds more meaning to the format in order to describe the classes and their properties in a more confined and conceptualized manner. Example could be: relationships between classes (e.g. whether the classes intersect or are disjoint), cardinality of the objects (e.g. "exactly 1"), equality of classes, characteristics of properties (e.g. transitivity), and many more. SPARQL stands for SPARQL Protocol and RDF Query Language. It is a query language for semantic web applications used to retrieve and work on the data stored in databases in RDF format [16]. Rule Interchange Format is the W3C standard

format used to define the rules used to exchange or share data on the web which is universal in nature. It is XML language which expresses rules for Web that can be executed by machines.

2.4.3 APPLICATIONS

The motive of Web 3.0 is to improve the functioning and usability of Web and the connected resources of Web with the help of Web 3.0 services. Some of the applications are:

- There are some servers that still expose some data systems which are in existence, using RDF format. Unstructured data can be converted into RDF format by using some techniques. One of the important data source could be relational databases as they are semi-structured. It reduces the cost of implementing it into the current system as it's server attaches itself easily to the existing system without affecting it's operations.
- There are some documents which are marked up some information in itself, an addition to HTML Meta tags, which are now-a-days used in web pages to give some information to search engines. So basically, these can be used as machine-understandable data about the data understandable by humans or it could totally be depicting metadata giving a collection of facts.
- There are some automated agents which can complete the tasks for the users.

Search engines or some data management system within an organization could make best use of above applications. Applications for Business purpose could be:

- Information coming from various sources can be integrated.
- Ambiguities can be eliminated due to a structured format.
- Retrieval of the information can be improved and so will information overload will be reduced.
- If given a domain, relevant information can be identified.

2.4.4 ONTOLOGIES

Historically, the term ontology has evolved from the metaphysics, the branch of philosophy, which basically deals with the nature of reality-of what exists. This branch analyses different types of existence with special attention to the relations between particulars and universals, between intrinsic and extrinsic properties, and between essence and existence. The traditional goal of ontological inquiry in particular is to divide the world "at its joints" to discover those fundamental categories or kinds into which the world's objects naturally fall. During the second half of the 20th century, philosophers extensively debated the possible methods or approaches to building ontologies without actually building any very elaborate ontologies themselves.

In the early 1990s, the widely cited Web page and paper "Toward Principles for the Design of Ontologies Used for Knowledge Sharing" by Tom Gruber is credited with a deliberate definition of ontology as a technical term in computer science. Gruber introduced the term to mean a specification of a conceptualization:

"Ontologies are often equated with taxonomic hierarchies of classes, class definitions, and the subsumption relation, but ontologies need not be limited to these forms. Ontologies are also not limited to conservative definitions — that is, definitions in the traditional logic sense that only introduce terminology and do not add any knowledge about the world. To specify a conceptualization, one needs to state axioms that do constrain the possible interpretations for the defined terms [17]."

2.4.5 COMPONENTS OF ONTOLOGIES

Ontologies generally have various structural similarities in spite of the different domains in which these are expressed. Most ontologies define classes (concepts), individuals (instances), their attributes and relations between them. The common components of ontology are:

- *Classes:* sets, collections, concepts, classes in programming, types of objects, or kinds of things.
- *Individuals*: instances or objects (the basic or "ground level" objects)
- *Attributes:* aspects, properties, features, characteristics, or parameters that objects (and classes) can have.

- *Relations:* ways in which classes and individuals can be related to one another.
- *Function terms*: complex structures formed from certain relations that can be used in place of an individual term in a statement
- *Restrictions*: formally stated descriptions of what must be true in order for some assertion to be accepted as input
- *Rules:* statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form
- *Axioms:* assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic.
- *Events:* the changing of attributes or relations

Ontology is generally encoded with the help of ontology languages.

2.4.6 TYPES OF ONTOLOGY

- *Domain ontology:* A domain ontology describes the concepts which are more of like real world entities. It provides meanings of the terms or concepts within the domain. For example, the word apple has several different meanings. If ontology is built for the domain of fruits, it would give meaning to “apple” word as a fruit but if ontology for the domain of smartphones brand is built, then it would model “apple” as the name of a company who manufactures smartphones. Since these domain ontologies define the concepts very specifically, they often become incompatible because some systems may need some sort of expansion and it will require the merging of some domain ontologies into more integrated and general representation. Use of different ontology languages, different purposes of developing ontology or dissimilar viewpoints of the users for a domain are some causes that leads to development of different ontologies. It is a very time consuming and expensive process to merge those ontologies that does not share same foundation as it is a manual process and no automatic technique exist till date. However, those domain ontologies that share same foundation basis can be automatically merged using some in-built plugins available in tools.

- *Upper ontology:* An upper ontology basically describes the model of some common entities and objects that generally relate to a wide range of the domain ontologies. It generally uses basic vocabulary which contains all the terms and their associated meanings as they are used in different contexts.
- *Hybrid ontology:* It is a combination of domain ontologies and upper ontologies.

2.5 LIGHT ON THE DOMAIN SELECTED

Ontology basically represents the concepts, their attributes and their inter-relationships about a domain. In this work, the domain is Ministries of Government of India. The reason of choosing this domain is due to the initiative recently proposed by Government of India which is "Digital India". This initiative is a big step to transform the country into a digitally empowered knowledge economy & includes projects that aim to ensure that government services are available to citizens electronically and people get benefit of the latest information and communication technology [18]. So, it would be efficient to develop an ontology representing all the departments of different ministries and their respective functions as it would conceptualize the structure of government, provide a unified framework for all the services provided and help respective ministries to collaborate with each other for policies, schemes and shared responsibilities in a homogeneous and global manner. This is basically an important feature of semantic web to share the information across the web with some predefined protocols and structure. Developing an ontology is more like identifying and specifying a set of data and its framework in order to use it in any other application. Problem-solving approaches, software agents and domain-independent applications make use of ontologies and knowledge retrieved from ontologies as data. The intent of building this ontology is also a preliminary step for a grander application of implementing sentiment analysis to apprehend government practices, policies, rules and monitoring performance. Hence, the idea is to envision "digital governance" by virtue of social web adoption, where government can take advantage of social platforms involving huge user participation.

2.6 NAÏVE BAYES CLASSIFIER

Naïve Bayes Classifier belongs to the family of simple probabilistic classifiers and it is based on Bayes' theorem. It assumes that the features are strongly independent. Naïve Bayes method is used since the 1950s and is a popular basic method to categorize the text. It uses frequencies of words in the document as the feature for classification. This method is still the one of the best methods of text classification among some advanced techniques like support vector machines etc. [19]. Naïve Bayes has a very crucial role on automatic medical diagnosis [20]. The advantage of Naïve Bayes classifier is its high scalability as it requires parameters in proportion to number of features or variable and that too linearly.

Naïve Bayes is a baseline and easy method to construct classifier which is a structure/system that associates some class values to objects or instances, which are Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, denoted as feature vector (in form of matrix). This is not just one algorithm but a group of algorithms sharing a common characteristic: all of the Naïve Bayes classifiers makes an assumption that the different features are independent of each other, given a domain.

As there are various kinds of probability models, it is very efficient to train Naïve Bayes classifier using some supervised learning techniques. In addition to this, sometimes Naïve Bayes classifiers makes use of maximum likelihood for estimating the parameters. In other words, Naïve Bayes model can be used by anyone even without even agreeing on Bayesian probability theorem or any other Bayesian methods.

In spite of the fact that Naïve Bayes classifier has very naïve structure and some simplified assumptions, this is being used widely and is giving efficient results in some real world situations. It has an advantage that only small set of data is required for training the classifier or for estimating the parameters needed for classification.

2.6.1 PROBABILISTIC MODEL

Conceptually, Naïve Bayes Model is a conditional probability model. If a problem instance given for classification, is denoted by a vector $x= (x_1, x_2, x_3, \ldots\ldots, x_n)$ where x_i represents i^{th} independent variable, the model gives some probabilities to these instances:

$$p(C_k|\ x_1, \ldots, x_n) \quad (1)$$

where k is the total number of possible classes.

If the number of features is large or the feature have a larger range for values, then using this model for assigning probability will become difficult. To deal with this problem, the probability defined above can be reformulated as:

$$p(C_k|x) = \frac{p(C_k)\ p(x|C_k)}{p(x)} \quad (2)$$

Generally, numerator in the above fraction is of interest because the denominator part is not dependent on C.

So, $$p(\ C_k|x_1, \ldots, x_n) \propto p(C_k)p(x_1, \ldots\ldots, x_n) \quad (3)$$

Now, Naïve Bayes model assumes that the conditional probabilities of all the variables are independent, so these terms can be reformulated as:

$$p(X|C_k) \propto \prod_{i=1}^{n} p(x_i|C_k) \quad (4)$$

Therefore,

$$p(C_k|X) \propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (5)$$

Using above formula, the case X will be labelled to the class C_k for which the probability $p(C_k|X)$ is maximum.

It is generally assumed that the continuous dataset are generally distributed along the Gaussian distribution. Let there be a set of training data which has an attribute, x which is continuous in nature. According to the Gaussian distribution, dataset is first divided according to the classes and then mean and variance of x is computed. Let for all the values of x in some class c, their mean be μ_c and that variance be σ_c^2. So the value of $p(x_i \mid C_k)$ for a particular x_i that belongs to set x, can be calculated as:

$$p(x_i|c) = \frac{1}{\sqrt{2\pi{\sigma_c}^2}}\, e^{-\frac{(x_i-\mu_c)^2}{2{\sigma_c}^2}} \quad (6)$$

2.7 RELATED WORK

In the initial work, researchers have used conventional methods to perform sentiment analysis on micro-blogging posts. There were two main approaches which were used to identify whether the text (a sentence, a paragraph or a document) possess positive, negative or neutral opinion. These are machine learning approach and lexicon based approach. The former approach of using machine learning techniques [21] includes training of sentiment classifier, in order to classify the textual corpora into positive, negative or neutral classes. Generally, unigrams and bigrams (n-grams of size 1 and 2 respectively) are used in training the classifiers [22]. The disadvantage of using machine learning methods is that it requires manual labelling of huge amount of text in order to train the classifier. Also, the labelling has to be done separately for each of the distinct domain to optimize the results for the given domain [23]. The second approach [24] [25] [26] includes use of some common opinion words which expresses sentiment (positive or negative). These words constitutes a dictionary which is called opinion lexicon. For this approach, tweets are not considered as normal text because only 140 characters are allowed in a tweet which imposes limitation on the length of words and phrases. The wide use of some everyday abbreviations, expressions and emoticons especially jargon has made it very peculiar to use it in the model. An obvious solution is to add these expressions in terms of words to the database but this would give ambiguous results because these words or expressions are generally dynamic and, changes constantly and frequently due to the changing trends on the Web. One more difficulty is that these jargon words are dependent on domain. So, when the lexicon based techniques are implemented on some informal text, it results in low recall value. As surveyed by [27], the following approaches can be used to improve the performance of sentiment analysis on Twitter posts as compared to the traditional methods:

1. For the dataset with noisy labels and distant supervision [28], emoticon vocabularies, as an example, can be created to represent sentiment and some supervised classifiers like Naïve Bayes (NB), Support Vector Machines (SVMs) and Maximum Entropy (MaxEnt) can be used for training purpose [29] [30].
2. A combination of machine learning techniques and feature engineering that is feature based model and tree based model, and also n-grams and lexicon characteristics) can be applied to improve the performance of sentiment analysis [31] [32].

There already exists some sentiment analysis techniques which use ontology based methods but none of the existing works deploys the ontology the way we have proposed the framework in this work. Some of the existing work is summarized below.

In [33], the authors present a method to populate an existing ontology for earthquake evacuation with tweets instances. In the proposed approach, some information related to earthquake like evacuation centres, product offered at the centres, time at which the tweet was posted are extracted. Also, some additional information like evacuation centre address (using Google Maps) can be retrieved from Web and is appended to the above information. Although the tweets do not contain this information, but it is acquired in real time. There are some other research works which includes building of ontology to represent micro-blog posts and some relations between users who use social networking service. This work, is however not related to the research work we have proposed.

In [1], Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades and Nick Bassiliades have proposed an approach of using ontologies for representing data and for extracting the features based on which tweets are extracted for sentiment analysis of smartphone reviews. The authors have used web service, OpenDover, to classify the tweets with sentiment grade. In [34], authors have proposed the approach of identifying only the negative comments related to United States Postal Service and performing sentiment analysis on those negative comments. Ontology has been deployed in order to identify the problem area e.g. the deliver type- mail or letter etc. Also, the negative sentiments have been analysed using SentiStrength tool. But due

to the use of tool which is no up to date according to the domain does not give expected results as claimed by authors.

In [35], authors have come up with an approach of building an ontology in form of sentiment ontology tree which is in terms of hierarchy defining positive side and negative side. An algorithm has been defined to identify sentiments according to the threshold values and weights of the words in the product reviews. But, there is no automatic process of getting the attributes which vary according to the reviews. In [36], the authors have extracted the reviews from Websites (basically Amazon.com) and then pre-processed using WordNet lexicon based method. Then fuzzy logic has been used to build ontology from the relationships found in the reviews. Using NLP rules, ontology and lexicons, overall polarity had been given to a product based on all the reviews.

In [37], the authors have proposed an approach to analyse the sentiments of movie reviews. It uses Natural Language Processing techniques to tag the data and then, build the domain ontology to extract the features based on which the reviews are classified into positive, negative or neutral. In [38], the authors present methodology in order to classify the product reviews of Chinese products. The reviews are first pre-processed and performed POS Tagging. Then, fuzzy domain ontology is built by identifying the relationship between features and the reviews. Similarity between sentiment orientation words in reviews and a set of sentiment words in fuzzy domain ontology is calculated and accordingly polarity is given.

In [39], the authors have presented a methodology for analysing sentiments of Portuguese movie reviews and hotel reviews. It uses domain specific ontology for feature extraction and after tagging of the words, SentiWordNet is used to classify the reviews.

To our best knowledge, there is no work related to text classification of Twitter posts for Ministries of India: their schemes, policies, rules, regulations etc. As the initiative of "Digital India" was recently proposed by Government of India, concept based text classification has not yet been performed till date.

CHAPTER 3: PROPOSED FRAMEWORK

Section 3.1 gives an overview of the research undertaken. Section 3.2 depicts the flowchart of the proposed work. Finally, section 3.3 describes the model in detail with the platforms used and code snippets.

3.1 PROPOSED APPROACH

The World Wide Web is a wide, hugely distributed, source of information which is like a never ending sea of knowledge. The web is growing at a rate which is unbelievable both in terms of users and content. The content is easily available on the Web and the users are willingly collaborating to the Web in terms of information, experiences, opinions, thoughts etc. Due to easy availability of content and active participation of users, some micro-blogging sites especially Twitter, have become a rich content for analysing the opinions of the users about their satisfaction/ dissatisfaction about any topic. In order to mine the opinions, opinion mining is intended to understand and analyse the opinions, and come up with some results or pattern which may help any organization, businesses or users. Towards "Digital India" initiative, in order to promote "Digital Governance", the policies, schemes, rules imposed by Government of India need to be monitored and tracked to see how these schemes are affecting people of India and if any amendments need to be done.

In response to the identified need to analyse the opinions of users related to the domain discussed above, we propose a novel hybrid approach involving integration of three things: application of Web 2.0 (i.e. Twitter), Semantic Web Features and Sentiment Analysis. This research seeks how semantic web can be used in sentiment analysis of Twitter posts. To uncover the opinion direction, we have defined the concepts based on which classification is done. The concepts relate to the schemes, policies, services provided by the Government of India to the people of India. The concepts described in a structured format, which has an advantage that it can be restructured without even disturbing rest of the process. These concepts will be used to extract the tweets and classify into one of the three classes: positive, negative or neutral. The main feature of our approach is the use of concept based techniques to extract

the tweets as it is making the process easier and automated. Using concepts provide a large set of topics which covers the need and increases the efficiency of overall system.

3.2 THE FRAMEWORK

The proposed work intends to accurately cover the topics which are defined by concepts and retrieve the relevant tweets from twitter. The tweets, after pre-processing are classified into positive, negative or neutral. Then the overall accuracy of the system is calculated based on percentage of correct output given by the model. The basic flowchart of the approach is:

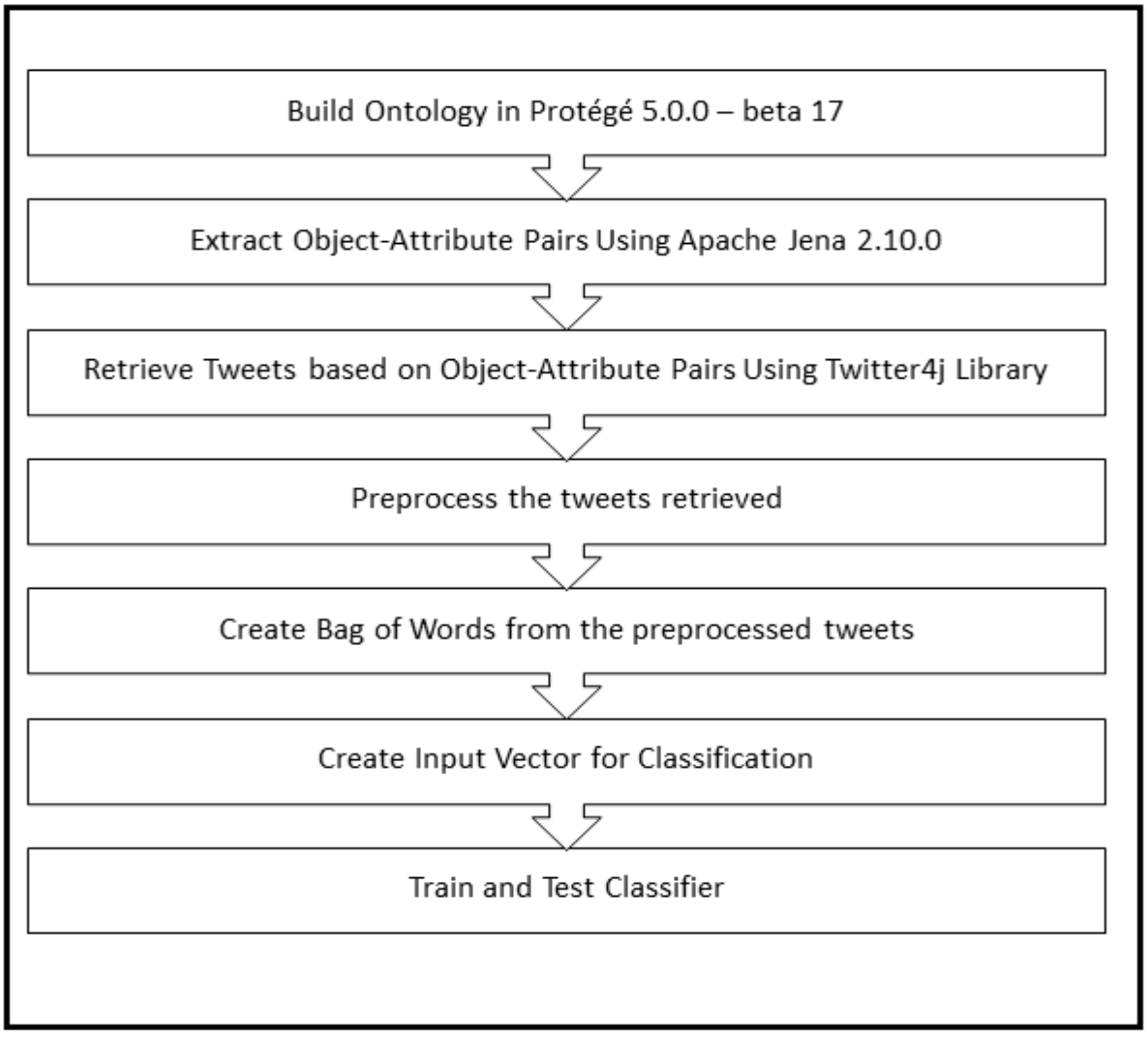


Figure 2

3.3 THE MODEL : SentIndiGov-O

After examining the research objectives and applications of Sentiment Analysis, we propose a model which fulfils the goals defined above. The detailed explanation is given below:

1. *Building Ontology*: As discussed earlier, the basic building block of semantic web is ontology. So, development of ontology is the preliminary step. The ontology of Ministries of Government of India, IndiGov-O, is developed in Protégé 5.0.0 – beta 17. It is a 4-level hierarchy which conceptualizes all the ministries of India, their departments and the functions of the departments. It includes 51 ministries, one independent department and one independent office [40]. This ontology is developed in Web Ontology Language (OWL). The editor used is Protégé 5.0.0 – beta 17. It is an open-source and free of cost tool or editor which can be used to build, develop or manipulate ontology. It offers a graphical user interface that helps in defining ontologies. There are some deductive classifiers which validates the consistency of ontology model and can also give inference after analysing the ontology. Protégé as an application is written in Java but the ontology it creates, is built in Web Ontology Language (OWL). Total number of registered users are over 200,000. It is one of the "leading ontological engineering editor". Stanford University and the University of Manchester has developed Protégé. The output or result can be integrated with any other compatible system to further build a more intelligent system. The ontology, IndiGov-O, developed is shown below:

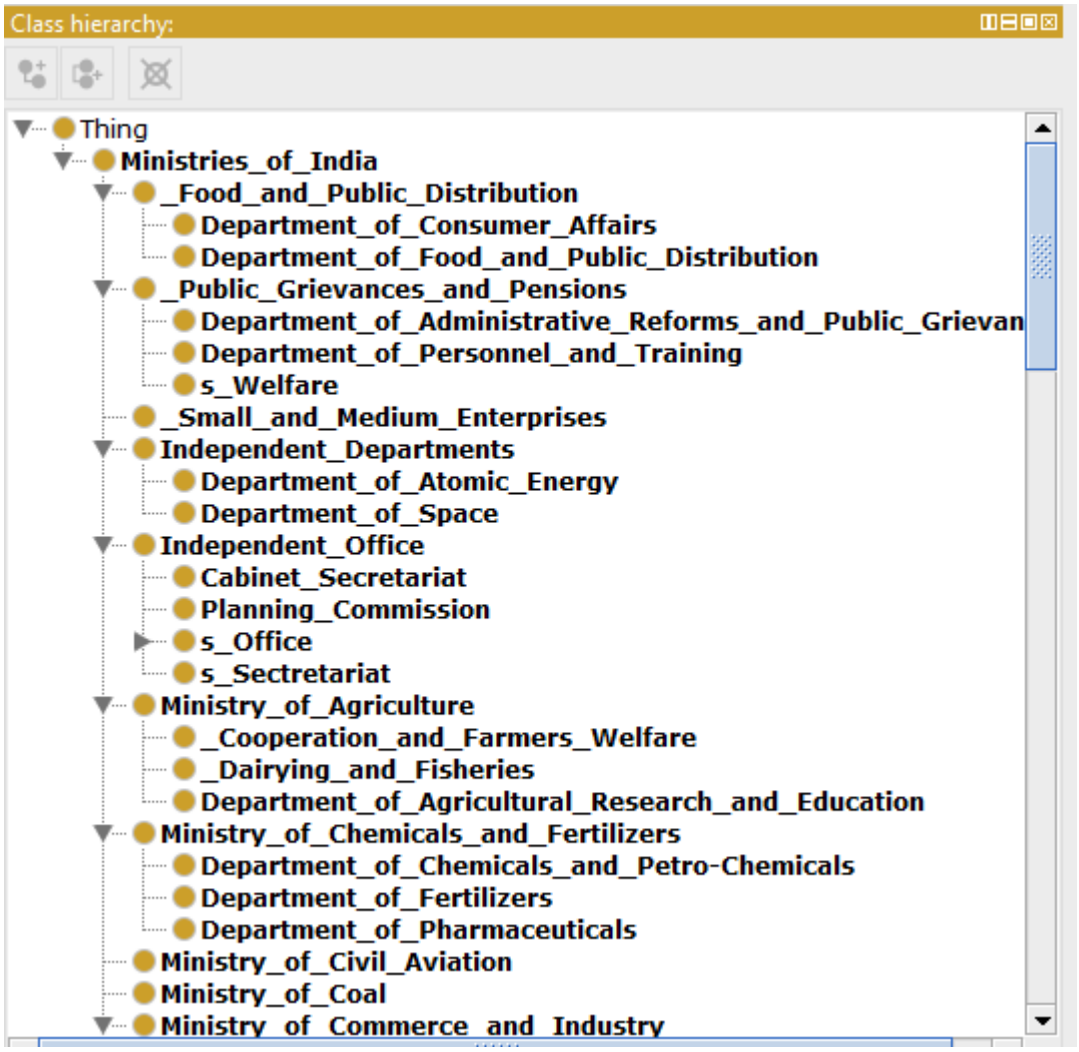


Figure 3

The above figure shows different ministries of Government of India and their departments.

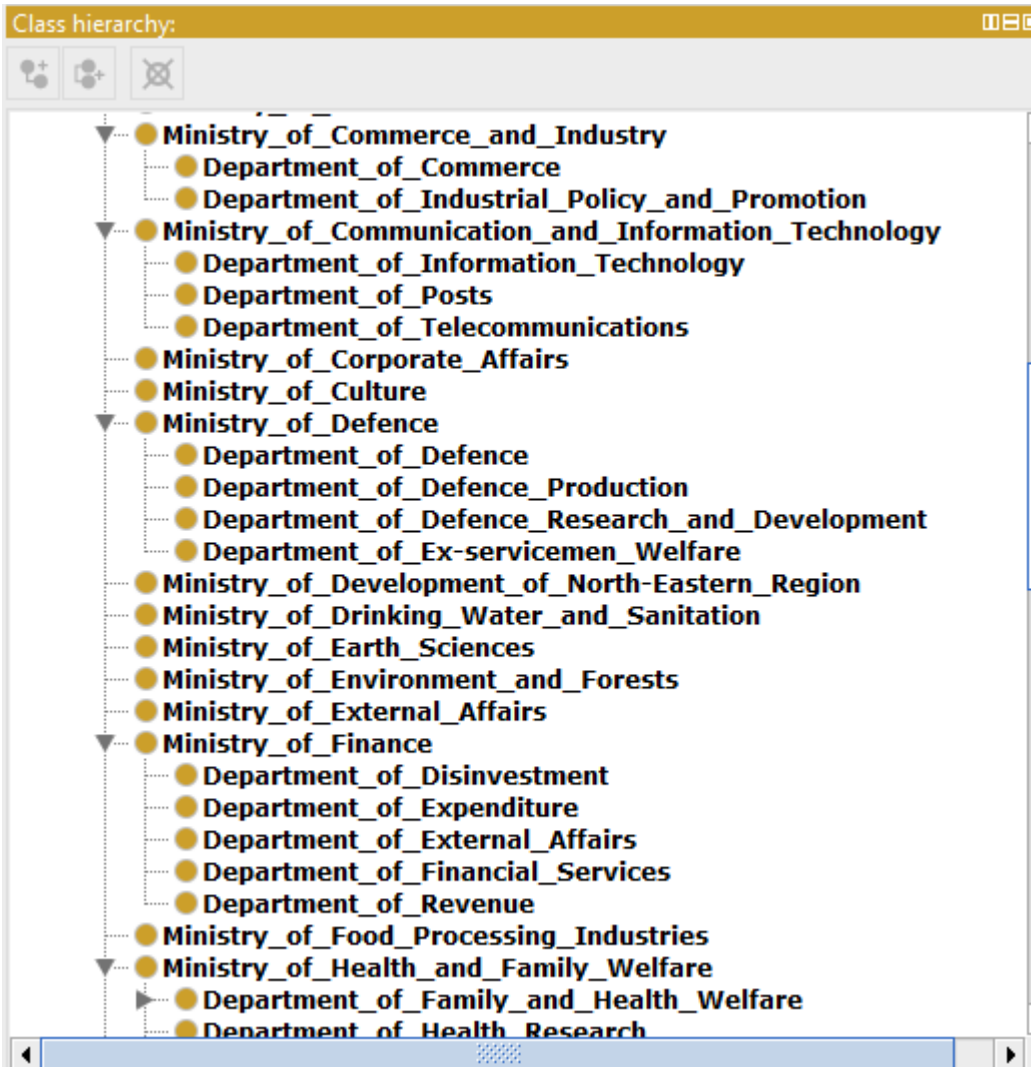

Class hierarchy:
Ministry_of_Commerce_and_Industry
Department_of_Commerce
Department_of_Industrial_Policy_and_Promotion
Ministry_of_Communication_and_Information_Technology
Department_of_Information_Technology
Department_of_Posts
Department_of_Telecommunications
Ministry_of_Corporate_Affairs
Ministry_of_Culture
Ministry_of_Defence
Department_of_Defence
Department_of_Defence_Production
Department_of_Defence_Research_and_Development
Department_of_Ex-servicemen_Welfare
Ministry_of_Development_of_North-Eastern_Region
Ministry_of_Drinking_Water_and_Sanitation
Ministry_of_Earth_Sciences
Ministry_of_Environment_and_Forests
Ministry_of_External_Affairs
Ministry_of_Finance
Department_of_Disinvestment
Department_of_Expenditure
Department_of_External_Affairs
Department_of_Financial_Services
Department_of_Revenue
Ministry_of_Food_Processing_Industries
Ministry_of_Health_and_Family_Welfare
Department_of_Family_and_Health_Welfare

Figure 4

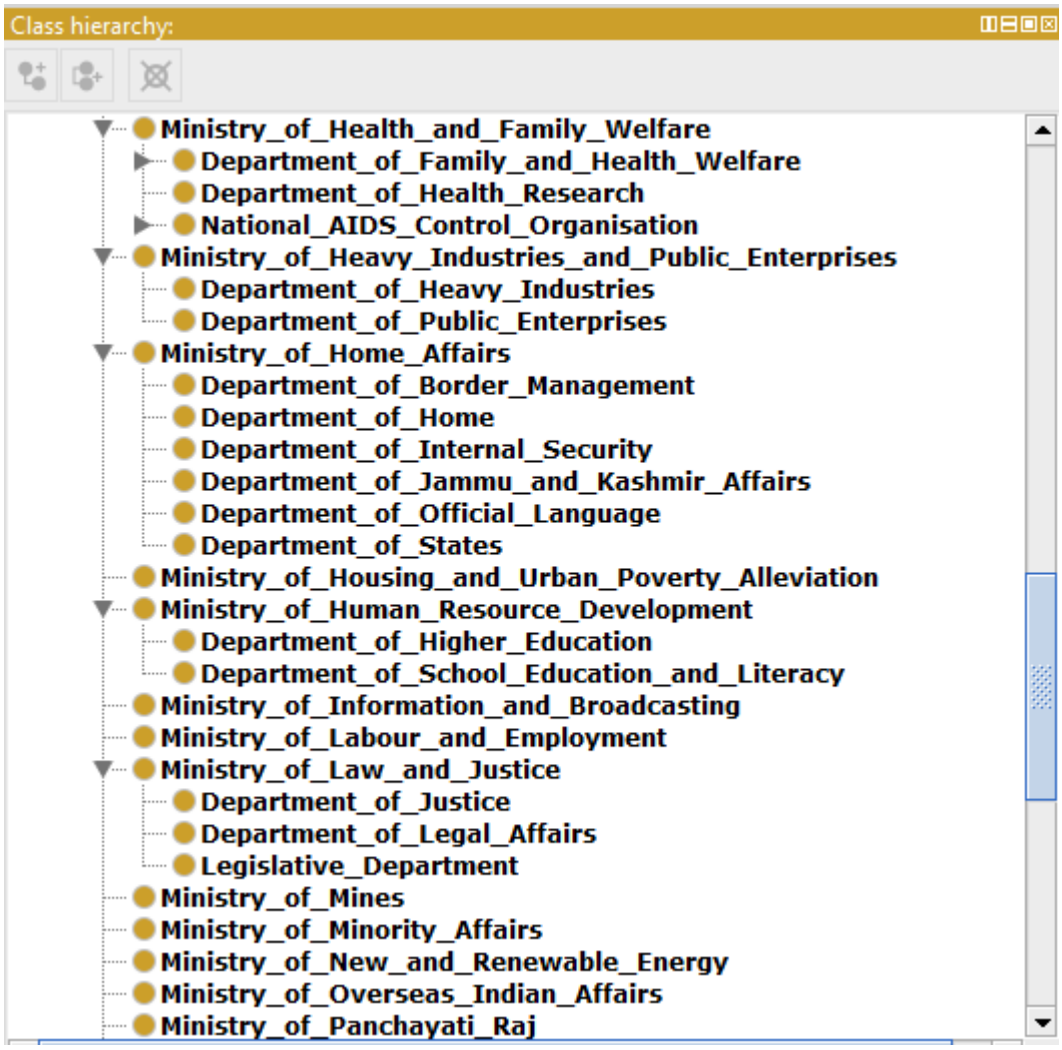

Class hierarchy:
Ministry_of_Health_and_Family_Welfare
Department_of_Family_and_Health_Welfare
Department_of_Health_Research
National_AIDS_Control_Organisation
Ministry_of_Heavy_Industries_and_Public_Enterprises
Department_of_Heavy_Industries
Department_of_Public_Enterprises
Ministry_of_Home_Affairs
Department_of_Border_Management
Department_of_Home
Department_of_Internal_Security
Department_of_Jammu_and_Kashmir_Affairs
Department_of_Official_Language
Department_of_States
Ministry_of_Housing_and_Urban_Poverty_Alleviation
Ministry_of_Human_Resource_Development
Department_of_Higher_Education
Department_of_School_Education_and_Literacy
Ministry_of_Information_and_Broadcasting
Ministry_of_Labour_and_Employment
Ministry_of_Law_and_Justice
Department_of_Justice
Department_of_Legal_Affairs
Legislative_Department
Ministry_of_Mines
Ministry_of_Minority_Affairs
Ministry_of_New_and_Renewable_Energy
Ministry_of_Overseas_Indian_Affairs
Ministry_of_Panchayati_Raj

Figure 5

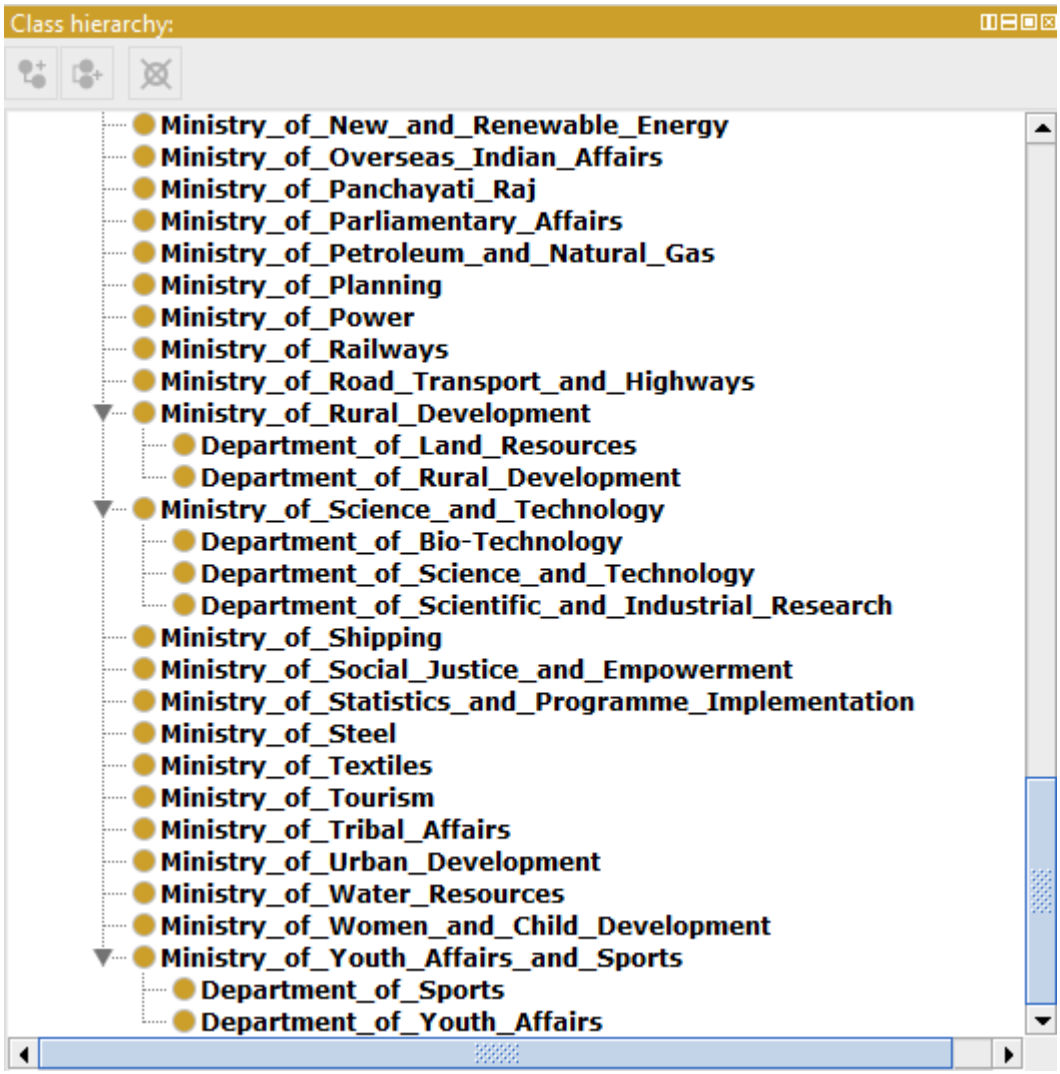

Class hierarchy:
Ministry_of_New_and_Renewable_Energy
Ministry_of_Overseas_Indian_Affairs
Ministry_of_Panchayati_Raj
Ministry_of_Parliamentary_Affairs
Ministry_of_Petroleum_and_Natural_Gas
Ministry_of_Planning
Ministry_of_Power
Ministry_of_Railways
Ministry_of_Road_Transport_and_Highways
Ministry_of_Rural_Development
Department_of_Land_Resources
Department_of_Rural_Development
Ministry_of_Science_and_Technology
Department_of_Bio-Technology
Department_of_Science_and_Technology
Department_of_Scientific_and_Industrial_Research
Ministry_of_Shipping
Ministry_of_Social_Justice_and_Empowerment
Ministry_of_Statistics_and_Programme_Implementation
Ministry_of_Steel
Ministry_of_Textiles
Ministry_of_Tourism
Ministry_of_Tribal_Affairs
Ministry_of_Urban_Development
Ministry_of_Water_Resources
Ministry_of_Women_and_Child_Development
Ministry_of_Youth_Affairs_and_Sports
Department_of_Sports
Department_of_Youth_Affairs

Figure 6

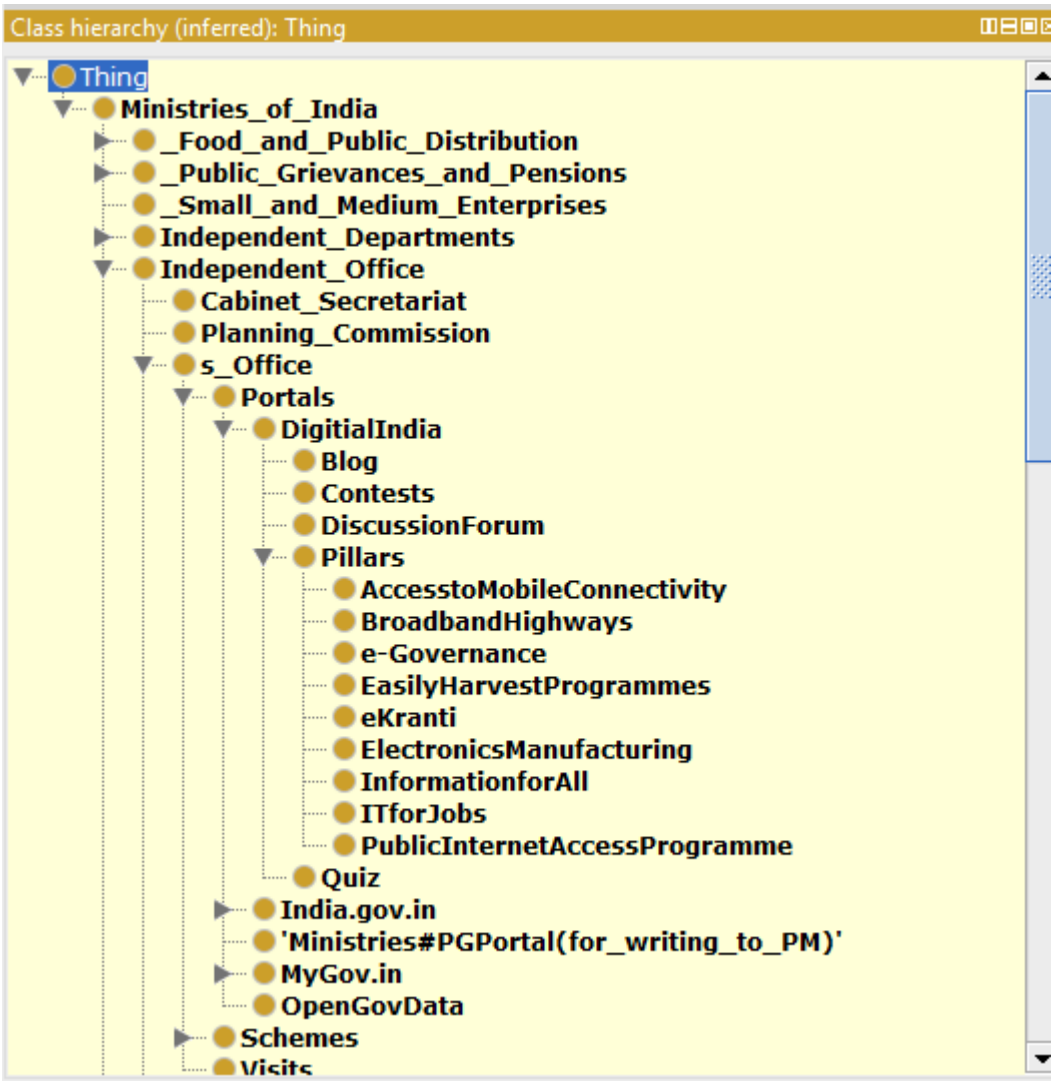


Figure 7

The above figure shows various schemes and functions of Prime Minister's Office under Independent Office. Above ontology is a 4-level hierarchy.

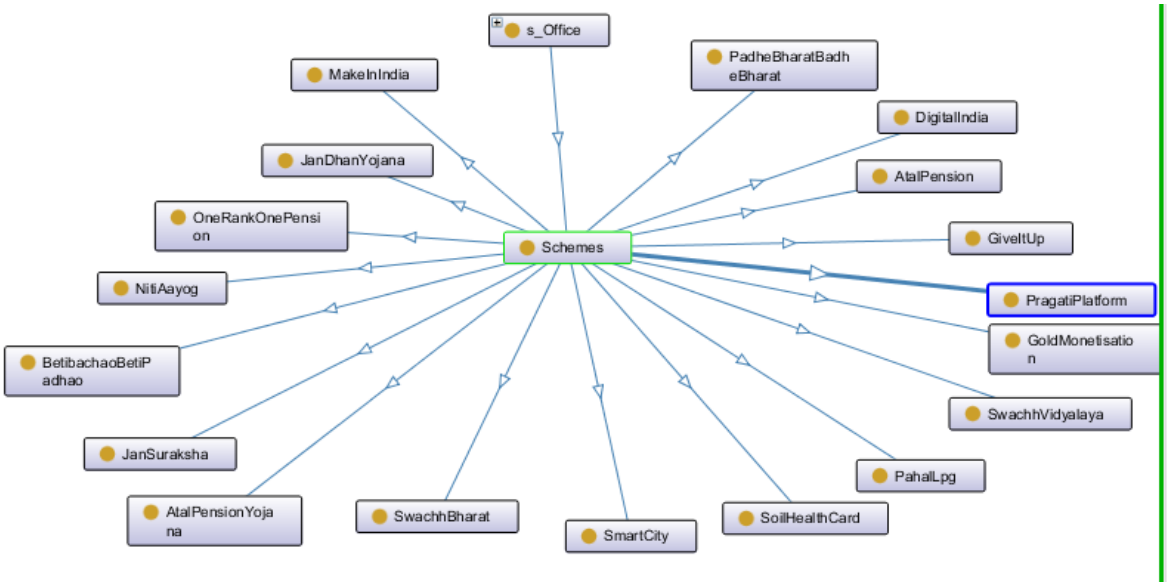


Figure 8

The above diagram shows the OntoGraf which shows the relationship between different entities. Schemes is the class and rest all entities are its subclasses.

2. *Extracting Object-Attribute Pairs*: This Web Ontology Language can be manipulated using Apache Jena Library. It is used to manipulate the data stored in RDF format. In order to perform sentiment analysis of Twitter posts, there is an obvious need to extract tweets from Twitter and in order to extract tweets, there is a requirement of some features on which related tweets are extracted. If performed manually, only 1 feature can be used at a time to retrieve tweets. But the use of Apache Jena and ontology easies the process of retrieval of tweets. It triggers the retrieval automatically using not just 1 but a list of features. So, these features in the form of object-attribute pair is extracted from ontology using Apache Jena 2.10.0 (These object – attribute pair is the class-subclass pair from IndiGov-O). It is an open source framework for Semantic Web applications. It is used as an application programming interface to extract or manipulate any data from the RDF schema. It can also be used to query the model using SPARQL engine. It also supports some internal reasoners for validation. The code snippet for this step is:

*ObjectAttributPair.java

```
import com.hp.hpl.jena.ontology.OntClass;
import com.hp.hpl.jena.ontology.OntModel;
import com.hp.hpl.jena.ontology.OntModelSpec;
import com.hp.hpl.jena.rdf.model.ModelFactory;
import com.hp.hpl.jena.util.FileManager;
import com.hp.hpl.jena.util.iterator.ExtendedIterator;

public class ObjectAttributPair {
    public static void main(String[] args) throws IOException {
        String filename = "Ministry.owl";
  OntModel inf = ModelFactory.createOntologyModel(OntModelSpec.OWL_MEM_MICRO_RULE_INF);
  InputStream inp = FileManager.get().open(filename);
  if (inp == null)
   throw new IllegalArgumentException("File: "+filename+" not found");
    inf.read(inp,"http://www.semanticweb.org/shadyinside/ontologies/Ministries");
    BufferedWriter writer =new BufferedWriter(new FileWriter("test.txt"));
    ExtendedIterator classes = inf.listClasses();
    while (classes.hasNext()) {
        OntClass Class = (OntClass) classes.next();

        for (Iterator i = Class.listSubClasses(); i.hasNext();) {

            System.out.println("Class: " + Class.getLocalName());
            OntClass c = (OntClass) i.next();
            if(!(Class.getLocalName().equals("Thing")) && !((c.getLocalName().equals("Nothing"))|| c
            {   writer.write(Class.getLocalName() + "   " );
            System.out.print("  Subclass: " + c.getLocalName() + "\n");
            writer.write(c.getLocalName() +"\n");
```

Figure 9

3. *Retrieval of tweets*: After the feature extraction, tweets are retrieved using Twitter4j library. It is a Java library for the Twitter API. Any Java application can be integrated with the Twitter service using Twitter4j library. It works using authentication keys provided by Twitter. It provides various methods to extract the tweets, to get user timeline, to get user-id etc. The code snippet for extracting tweets using Twitter4j library is:

SearchTweets.java

```
}
ConfigurationBuilder cb= new ConfigurationBuilder();
cb.setDebugEnabled(true).setOAuthConsumerKey("5NoQeJ7aMKcGV2Bm7GN41gf10");
cb.setDebugEnabled(true).setOAuthConsumerSecret("fOyL9BDZNPd7iRnKF9blhbEMkhOpw0ZUYdLaleOGsCKBFyaM6E

cb.setDebugEnabled(true).setOAuthAccessToken("2616851598-jcXbbBUtDfYNNS0WG4YslP6TzmakzOtj1enB8Bf");
cb.setDebugEnabled(true).setOAuthAccessTokenSecret("rJS1VRNXSFYjjLdiHLD0IABJzHDWaBqFT5gXTHFPZhNhf")
TwitterFactory tf=new TwitterFactory(cb.build());
twitter4j.Twitter twitter = tf.getInstance();
try {
    BufferedReader in = new BufferedReader( new FileReader("test.txt"));
    BufferedWriter writer =new BufferedWriter(new FileWriter("test2.txt"));
    String line= null;
    while((line=in.readLine())!= null)
    {
    String arr[]= line.split("  ");
    Query query = new Query("#"+arr[0]+"#"+arr[1]+"Government");
    QueryResult result;
    do {
        result = twitter.search(query);
        List<Status> tweets = result.getTweets();
        for (Status tweet : tweets) {
                System.out.println("@" + tweet.getUser().getScreenName() + " - " + tweet.getText())
                writer.write(tweet.getText()+"\n");
        }
    } while ((query = result.nextQuery()) != null);
    }
    in.close();
    writer.close();
    System.exit(0);
```

Figure 10

4. *Pre-processing the tweets*: The tweets retrieved in the previous process needs to be pre-processed for analysing the sentiments of these tweets. Pre-processing mainly includes removing unwanted things from the lot of tweets which will not do any positive contribution to the sentiment. Removing those which are neutral and does not play any significant role, makes the process of classification much easier and efficient. It includes:

    - Removing URL
    - Removing usernames
    - Removing additional whitespaces
    - Converting letters to lowercase
    - Replacing #word with word
    - Removing retweets
    - Deleting stop words
    - Stripping punctuation marks for feature vector.

```
input_file = 'newfinaltweet.txt'
output_file = 'newfinaltweet1.txt'
myfile= open(output_file,'a')
with open(input_file) as f:
    lines = f.readlines()
for line in lines:
    filtered_line = [w for w in line.split() if not w in stopWords]
    filtered_line = ' '.join(filtered_line) + '\n'
    myfile.write(filtered_line)
myfile.close()
```

```
def processTweet(tweet):
    # process the tweets
    #Convert to lower case
    tweet = tweet.lower()
    #Convert www.* or https?://* to URL
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','',tweet)
    #Convert @username to AT_USER
    tweet = re.sub('@[^\s]+','',tweet)
    #Remove additional white spaces
    tweet = re.sub('[\s]+', ' ', tweet)
    #Replace #word with word
    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
    #tweet= re.sub('[^\s]+', '', tweet)
    tweet = tweet.strip('\'"?,.')
    return tweet
```

Figure 11

5. *Creating Bag-of-Words*: After pre-processing the tweets, in order to create input vector for the classification process, a bag of words is created. Bag of words contains all the words

from the tweets. All the duplicated words are removed and so unique set of words is built.

```
fp = open('newtweets1.txt', 'r')
line = fp.readline()
tweetfile = open('newfinaltweet.txt','w')
st = open('stopwords.txt', 'r')
stopWords = getStopWordList('stopwords.txt')
featurevector = open('BagofWords.txt','w')
featureList = []
while line:
    processedTweet = processTweet(line)
    tweetfile.write(processedTweet)
    tweetfile.write("\n")
    featureVector = getFeatureVector(processedTweet)
    featureList.extend(featureVector)
    line = fp.readline()
#end loop
fp.close()
```

Figure 12

6. *Creating Input Vector*: After creating bag of words, a vector needs to be created which will serve as input to the classifier. This vector called as input vector depends on the bag of words and the individual tweets. Creation of input vector takes into account the number of the words which are common in each tweet and bag of words. The reason of having this input vector is because the classifier used in this work takes numerical data as input to classify the data.

```
arr = []
pre = open('newfinaltweet.txt','r')
outputfile = open('newoutputvector.csv', 'a')
redy = csv.writer(outputfile)
total=0
for line in pre:
    total += 1
    arr= line.split()
    arr1 = [0] * rows
    for i in range(len(arr)):
        if arr[i] in featureList:
            arr1[featureList.index(arr[i])] += 1
    redy.writerow(arr1)
print total
```

Figure 13

7. *Training and Testing*: The classifier used in here is Naïve Bayes Classifier. It uses the input vector as training data and classify the testing data into three classes i.e. positive, negative and neutral. Total 1019 tweets are used as training data and 503 tweets are used as testing data.

   The algorithm for training phase used in Naïve Bayes classifier is:

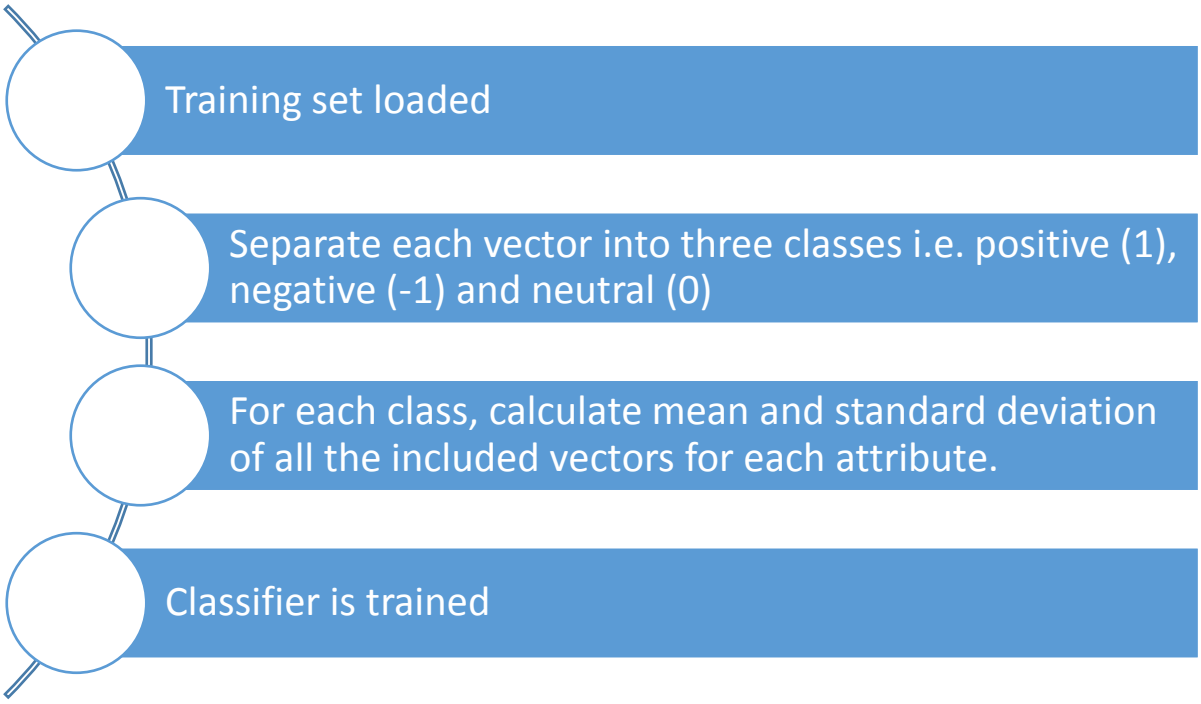


Figure 14

   The algorithm for testing phase in Naïve Bayes classifier is:

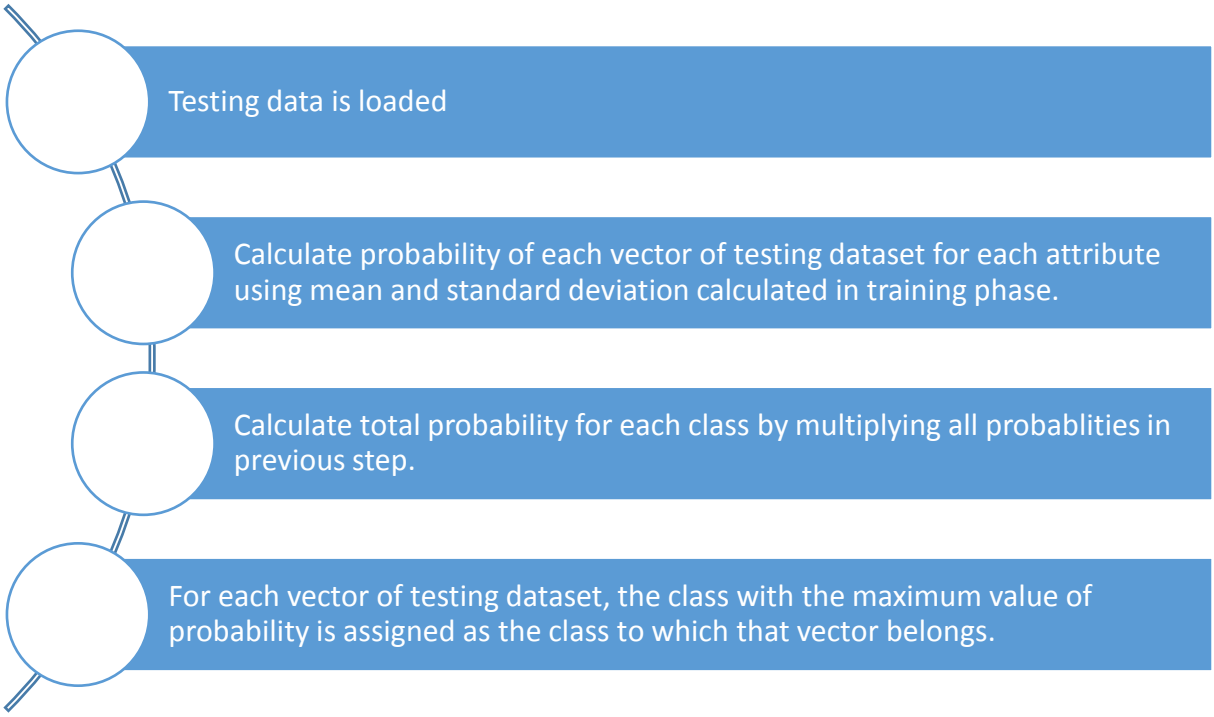


Figure 15

3.3.1 IMPLEMENTATION DETAILS OF NAÏVE BAYES:

Firstly, the input vector file in CSV format is loaded. Next, the dataset is split into training set and testing set. Training set is used to make the predictions and testing set is used to testing data is used to evaluate the accuracy of the model. The split ratio is generally 0.67. So the training data is 67% of the dataset and testing data is 33% of the dataset. Then the training data is separated by class values into positive (1), negative (-1) and neutral (0). This is done to calculate the statistics.

Now, these separated instances are summarized according to the class values. Mean of each attribute of each class is calculated. The mean is the central middle or central tendency of the data, so it is used as the middle of the Gaussian distribution when probabilities are calculated. Also standard deviation of each attribute for a class value is calculated. The standard deviation describes the variation of spread of the data, and it will be used to characterize the expected spread of each attribute when probabilities are calculated. The standard deviation is calculated as the square root of the variance. The variance is calculated

as the average of the squared differences for each attribute value from the mean. So, finally mean and standard deviation is calculated attribute wise as a summary of all the instances of a class.

Now, for each testing dataset instances, probability, according to Gaussian function, for each attribute value is calculated using mean and standard deviation calculated in the training phase. As the probabilities of all the attributes belonging to a class have been calculated, it can be combined to calculate an overall probability for a data instance for each class to make the prediction. The probability for each data instance is calculated by multiplying together the attribute probabilities for each class. Now, the instance belongs to that class which has the highest class probability calculated above. The predictions are compared to the class values in the test dataset and an overall accuracy is calculated.

CHAPTER 4: RESULTS

The results according to the steps defined in the proposed framework are:

4.1 BUILDING ONTOLOGY:

Below is OntoGraf showing various ministries of India.

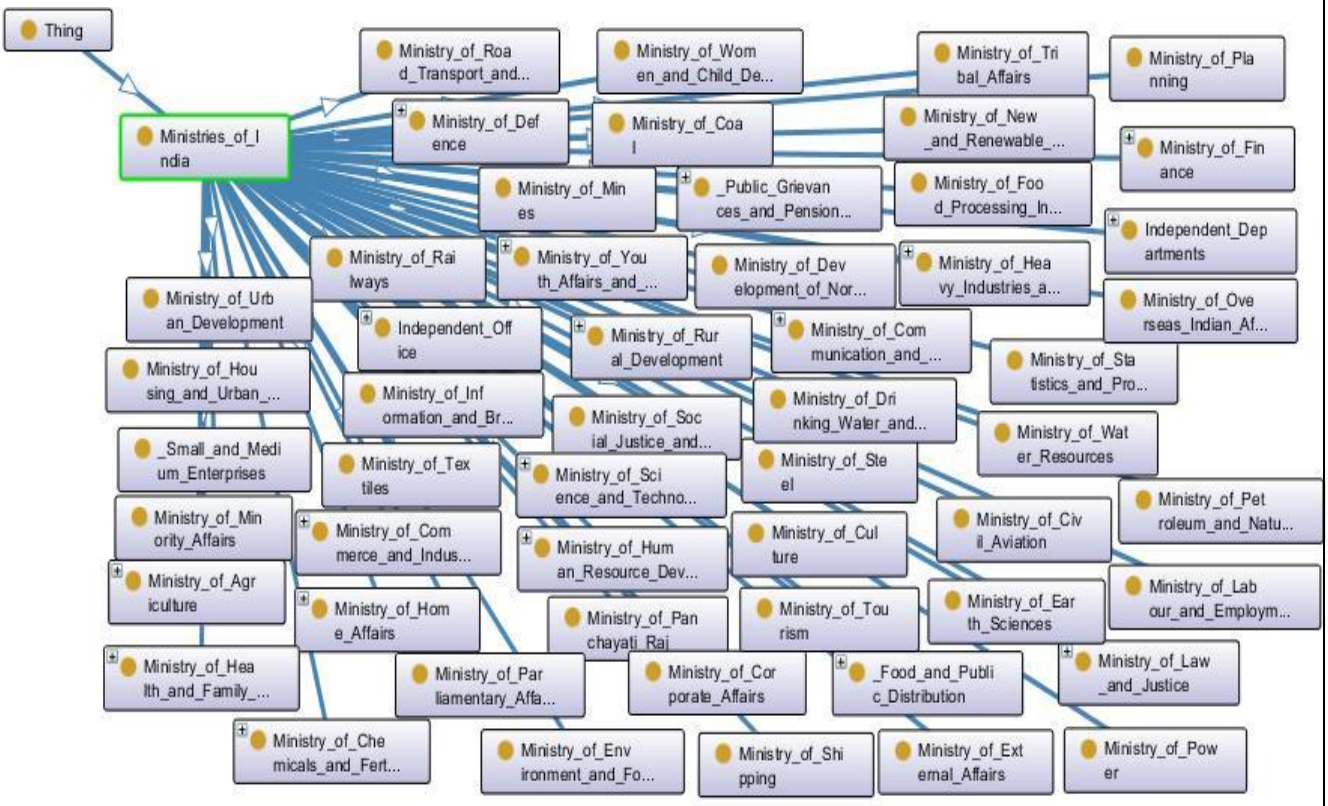


Figure 16

4.2 EXTRACTING OBJECT-ATTRIBUTE PAIRS:

The following image shows some of the object- attribute pairs extracted from the above ontology.

```
Portals  OpenGovData
Portals  MyGov.in
Portals  India.gov.in
Portals  DigitialIndia
Portals  DiscussionForum
Portals  Contests
Portals  Quiz
Portals  Pillars
Portals  e-Governance
Portals  eKranti
Portals  ITforJobs
Portals  BroadbandHighways
Portals  PublicInternetAccessProgramme
Portals  AccesstoMobileConnectivity
Portals  InformationforAll
Portals  EasilyHarvestProgrammes
Portals  ElectronicsManufacturing
Portals  Blog
Portals  GOIDirectory
Portals  e-governance
Portals  Webcasts
Portals  Calendar
Portals  OnlineServices
Portals  e-greetings
Portals  Poll
Portals  Groups
Portals  Discuss
Portals  Blogs
Portals  Tasks
Portals  MannKiBaat
Ministry_of_Finance  Department_of_Revenue
Ministry_of_Finance  Department_of_Financial_Services
Ministry_of_Finance  Department_of_External_Affairs
Ministry_of_Finance  Department_of_Expenditure
Ministry_of_Finance  Department_of_Disinvestment
Ministry_of_Defence  Department_of_Ex-servicemen_Welfare
```

Figure 17

4.3 RETRIEVING TWEETS BASED ON OBJECT-ATTRIBUTE PAIRS:

The following figure shows some of the tweets retrieved from Twitter using Twitter4j.

Figure 18

4.4 PRE-PROCESS THE TWEETS RETRIEVED:

The tweets retrieved above are not pre-processed yet. The following figure shows the tweets after pre-processing phase:

```
opendata default gov't. that's joined 50 supporting opengovdata act
opengovdata act requires federal agencies publish info online, using non-proprietary, opendata formats
learn opengovdata act? explore here: opendata
api election results (not us)? api opengov elections opengovdata
womentransform story-25: muskan, change age. support cause
ecomaterionics shud followed rational &amp amp recuperation resources frm water grains b4 late
read 25 finalists women transforming india! vote them
104-yr-old sold goats build toilet &amp;amp; inspired all. vote 4 kuwarbai: womentransform
spl drive cleanliness govt offices/ bldgs fr 16-31 may. photos uploaded mu
namodevelopingindia :direct benefit transfer (dbt) paradigm-reaching correct beneficiary
month's mannkibaat 22nd may.there multiple contribute
won't walk. athlete instead! vote 4 deepa malik: womentransform
womentransform story-14 phani trivedi aims zero waste! support vote
womentransform story-9 roshni 100 women achievers support vote
transformingindia: quantum jump fdi telecom sector years
prime minister transformingindia website - portal positive news:
womentransform story-5 moriam toppo,securing land rights. vote mygov via sms
womentransform story-6 lata mane,instilling hope tomorrow. vote
womentransform story-10 kuwarbai yadav, inspiration us. vote mygov
womentransform story-12 pavithra ys, enabling disabled. vote mygov sms
womentransform story-13 naheed aqueel, transforming lives. vote
womentransform story-18 indiritta,making quality healthcare accessible. click vote
transformingindia: features pmujjwalayojna launched pm modi-lpg bpl families.
transformingindia: parliament passes bill repeal additional 1053 redundant laws.
e-health laws regulations india businesses digital india project
growth technology business india impacted shortcomings digital india project urge
ekranti initiative surely hour! digital transformation certain
passport, pension dl edelivered ekranti aadhaarcard itr mumbai
digitalindia's ekranti offer passports pension online |
looking forward passport option existing ekranti
mother's day, share favourite e-greetings mother design win. click share
congratulations winners! results egreetings baisakhi contest announced. check name list. visit
fckfckshktimaan: ngoindia n4 paul foundation funding organisation 71195 ngo i am surprised u check too
check national portal india
ngoindia 1 according govt.india website 71195 ngo's registered partnership scheme
ngoindia 4 paul foundation funding organisation 71195 ngo i am surprised u check too
```

Figure 19

4.5 CREATING BAG OF WORDS:

The list of words created after pre-processing is shown below:

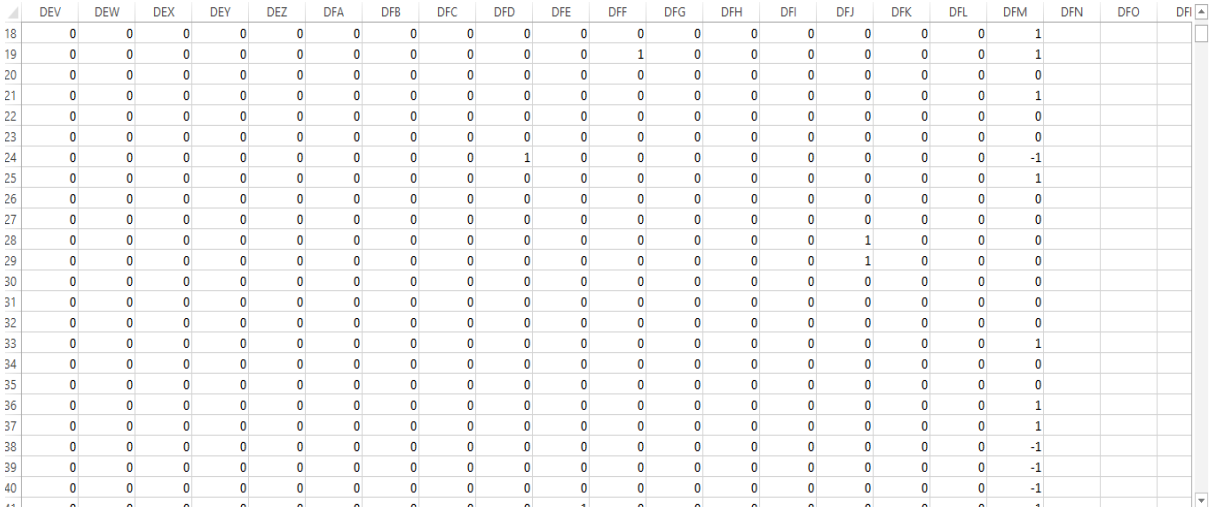
```
opendata
default
joined
supporting
opengovdata
opengovdata
act
requires
federal
agencies
publish
info
online
using
opendata
learn
opengovdata
act
explore
opendata
api
election
results
api
opengov
elections
opengovdata
womentransform
muskan
change
age
support
cause
ecomaterionics
shud
followed
```

Figure 20

Total number of attributes in bag is 2872.

4.6 CREATING INPUT VECTOR:

The input vector of the tweets thus created is shown below:



	DEV	DEW	DEX	DEY	DEZ	DFA	DFB	DFC	DFD	DFE	DFF	DFG	DFH	DFI	DFJ	DFK	DFL	DFM	DFN	DFO	DFI
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
19	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1			
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
24	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	-1			
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0			
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0			
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1			
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1			
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1			

Figure 21

The last column in above matrix is the sentiment value/ class value given to each tweet (row in above matrix).

4.7 TRAINING AND TESTING CLASSIFIER:

The following table shows some sample tweets with expected values of sentiments and the predicted values of sentiments as outputs from the classifier.

S. NO.	TWEETS	EXPECTED VALUE	PREDICTED VALUE
1.	school agarwal vidhya vihar surat guj school launched beti bachao beti padhao providing free education daughters .	1	1

2.	programs beti bachao beti padhao aimed girls india irrespective caste & comm…	1	1
3.	women child development director renu phulia 'beti bachao beti padhao': via	0	0
4.	transformingindia: check infographics pmsurakshabimayojana	0	0
5.	railways ;one iron chair signal lamp record book & loneliness journey ...pathetic	-1	-1
6.	modi govt becoming victim digital vertigo disconnected real world imprisoned virtual reality!	-1	-1
7.	afternoon ajubaa! read what's digital india. ajubaa makeindigitalindia	1	1
8.	seelampur: india's digital underbelly phones die -	-1	0
9.	1.55 lakh pregnant women extended financial assistance janani suraksha yojana 2yrsofgatisheelgujarat gujarat	1	0
10.	national health mission jish se hume job milta tha wo close ho gaya plzz modiji hume job chahiye homoeopathy dr. job	-1	0

Table 1

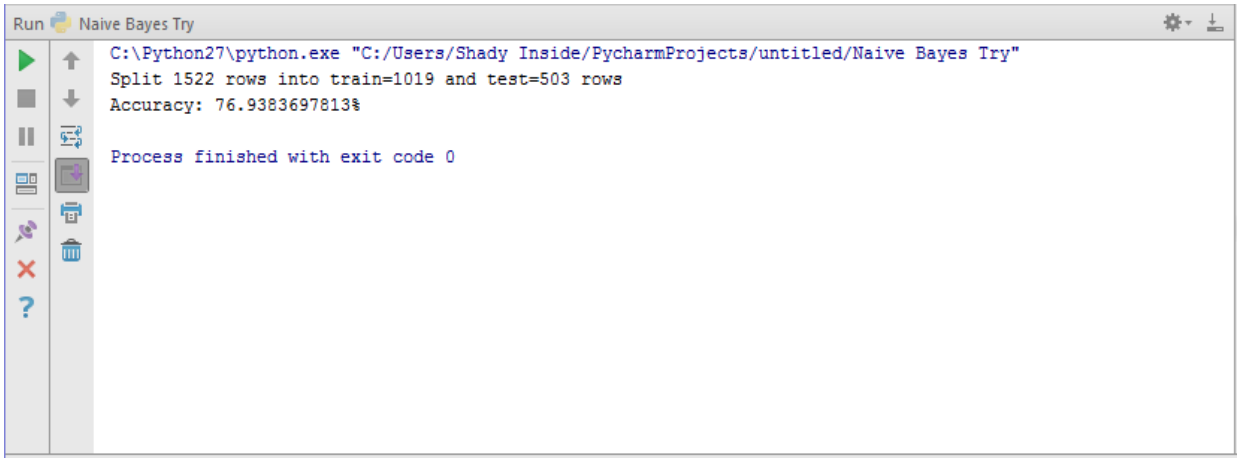

Run
Naive Bayes Try
C:\Python27\python.exe "C:/Users/Shady Inside/PycharmProjects/untitled/Naive Bayes Try"
Split 1522 rows into train=1019 and test=503 rows
Accuracy: 76.9383697813%
Process finished with exit code 0

Figure 22

CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

Since the rise of Web 2.0 and its related services and technologies like social networking sites, wikis and blogs, sentiment analysis has become one of the rapidly growing research area. The recent extensive use of micro-blogging services, especially Twitter, has given a lot of attention to opinion mining of micro-blogging posts. There are various machine learning techniques used in performing opinion mining on tweets with a disadvantage of treating each tweet as a uniform sentence and assigning a score according to sentiment of the post. This work proposes an approach of deploying ontology based techniques to in order to determine the subjects/ topics discussed in the tweets, which are related to Ministries of India: their policies, rules, schemes, and analysing the sentiment (positive, negative and neutral) of tweets retrieved (based on the concepts defined in Ontology). The results have been shown by using the Naïve Bayes model for the classification task, which shows that using ontology automates the process of retrieving tweets on various topics (and not just one) and then classifying it accordingly. Thus it also shows the integration of features of Semantic web, application of Web 2.0 (Twitter) and Sentiment analysis which leads to an automated overall process.

5.2 FUTURE WORK

The future improvements in the proposed work could be the use of various different machine learning techniques for classification process of tweets retrieved like Support Vector Machines (SVMs), Maximum Entropy (MaxEnt) or use of any neural networks like Probabilistic Neural Network, Recurrent Neural Network to optimize the performance of the model. Also, if analysis is needed for all the topics but separately like to monitor the performance of each scheme initiated by Government of India, then some changes in the modules of the system can be made accordingly.

REFERENCES

1. Kontopoulos, Efstratios, et al. "Ontology-based sentiment analysis of twitter posts." Expert systems with applications 40.10 (2013): 4065-4074.
2. http://www.digitalindia.gov.in/
3. Kumar, Akshi, and Teeja Mary Sebastian. "Sentiment analysis on twitter." IJCSI International Journal of Computer Science Issues 9.4 (2012): 372.
4. Kumar, Akshi, and Mary Sebastian Teeja. "Sentiment analysis: A perspective on its past, present and future." International Journal of Intelligent Systems and Applications 4.10 (2012): 1.
5. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", Proceedings of EMNLP, 2002, pp. 79—86.
6. Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, (2002), Philadelphia, Pennsylvania, 417-424.
7. Bo Pang and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", Proceedings of ACL, 2005, pp. 115—124.
8. Benjamin Snyder, Regina Barzilay, "Multiple Aspect Ranking using the Good Grief Algorithm", Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL), 2007, pp. 300–307.
9. Vasilis Vryniotis, "The importance of Neutral Class in Sentiment Analysis", 2013.
10. Moshe Koppel, Jonathan Schler, "The Importance of Neutral Examples for Learning Sentiment", Computational Intelligence 22, 2006, pp. 100–109.
11. Maite Taboada, Brooke, Julian, "Lexicon-based methods for sentiment analysis", Computational Linguistics, Volume 37 Issue 2, June 2011, pp. 272–274.
12. Bo Pang, Lillian Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval 2(1-2), 2008.

13. Mihalcea, Rada, Carmen Banea, and Janyce M. Wiebe. “Learning multilingual subjective language via cross-lingual projections.” 2007.
14. Su, Fangzhong, and Katja Markert. "From words to senses: a case study of subjectivity recognition." Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008.
15. Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.
16. Seaborne A, Harris S. SPARQL 1.1 Query Language. W3C Recommendation; 2013.http://www.w3.org/TR/2013/REC-sparql11-query-20130321/.
17. Gruber T R. “A Translation Approach to Portable Ontology Specifications”, Knowledge Acquisition; 1993, Vol. 5 No. 2, p.199–220.
18. Digital India Programme http://www.digitalindia.gov.in/content/about-programme.
19. J. Rennie, L. Shih, J. Teevan, D. Karger. “Tackling the poor assumptions of Naive Bayes classifiers”. ICML, 2003.
20. Irina Rish, “An empirical study of the naive Bayes classifier (PDF)”, IJCAI Workshop on Empirical Methods in AI, 2001.
21. B. Pang & L. Lee “Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval”, 2008, 2(1–2), 1–135.
22. D. De Kok & H. Brouwer, “Natural language processing for the working programmer”, 2012.
23. A. Aue & M. Gamon “Customizing sentiment classifiers to new domains: a case study”, in Proceedings of the international conference on recent advances in natural language processing (RANLP-05), 2005, pp. 207–218.
24. N. Kaji & M. Kitsuregawa, “Building lexicon for sentiment analysis from massive collection of HTML documents”, In Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 1075–1083.

25. A. Neviarouskaya, H. Prendinger & M. Ishizuka, “SentiFul: Generating a reliable lexicon for sentiment analysis”. In Proceedings of the affective computing and intelligent interaction and workshops (ACII), 3rd international conference on affective computing and intelligent interaction and workshops, IEEE, 2009, (pp. 10–12).
26. M. Taboada, J. Brooke, M. Tofiloski, K. Voll & M. Stede, “Lexicon-based methods for sentiment analysis”. Computational Linguistics, 2011, pp. 267–307.
27. H. Saif, Y. He & H. Alani, “Alleviating data sparsity for twitter sentiment analysis”. In 2nd Workshop on making sense of microposts (#MSM2012): Big things come in small packages at World Wide Web (WWW), 2012, pp. 2–9.
28. J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification”. In Proceedings of the ACL-05, 43nd meeting of the association for computational linguistics, Association for Computational Linguistics, 2005.
29. A. Pak & P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining”. In Proceedings of the 7th international conference language resources and evaluation (LREC ‘10), European Language Resources Association, ELRA, 2010.
30. L. Barbosa & J. Feng, “Robust sentiment detection on twitter from biased and noisy data”, In Proceedings of the 23rd international conference on computational linguistics: Posters (COLING ‘10), 2010, pp. 36–44.
31. A. Agarwal, B. Xie, I. Vovsha, O. Rambow & R. Passonneau, “Sentiment analysis of twitter data”, In Proceedings of the ACL 2011 workshop on languages in social media, 2011, pp. 30–38.
32. E. Kouloumpis, T. Wilson & J. Moore, “Twitter sentiment analysis: The good the bad and the OMG!”, In Proceedings of the ICWSM, 2011.
33. I. Iwanaga, T. M. Nguyen, T. Kawamura, H. Nakagawa, Y. Tahara & A. Ohsuga, “Building an earthquake evacuation ontology from twitter”, In Proceedings of the IEEE international conference on granular computing (GrC), 2011, pp. 306–311.
34. Pratik Thakora, Dr. Sreela Sasib, “Ontology-based Sentiment Analysis Process for Social Media Content”, INNS Conference on Big Data, Elsevier, 2015.

35. Wei Wei, Jon Atle Gulla, "Sentiment Learning on Product Reviews via Sentiment Ontology Tree", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pages 404–413.
36. Raymond Y. K. Lau, Chapmann C.L. Lai, jian Ma, Yuefeng Li, "Automatic Domain Ontology Extraction For Context-Sensitive Opinion Mining", International Conference on Information Systems, (ICIS), 2009.
37. Khin Phyu Phyu Shein, "Ontology based combined approach for Sentiment Classification", Proceedings Of The 3rd International Conference On Communications And Information Technology, 2009.
38. Hanshi Wang, Xinhui Nie, Lizhen Liu and Jingli Lu, "A Fuzzy Domain Sentiment Ontology based Opinion Mining Approach for Chinese Online Product Reviews", Journal Of Computers, Vol. 8, No. 9, September 2013.
39. Larissa A. Freitas and Vieira Renata, "Ontology based feature level opinion mining for portuguese reviews", Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013.
40. Ministries of India http://goidirectory.nic.in/union_categories.php?ct=E002.

APPENDIX A

List of Publications

Accepted:

1. A. Kumar, A. Sharma and A. Joshi, *IndiGov-O: An ontology of Indian Government to empower Digital Governance*, India International Conference on Information Processing (IEEE), 2016.

Communicated:

1. A. Kumar and A. Joshi, Ontology Based Tool for Sentiment Analysis to empower Digital Governance, International Conference on Computational Intelligence in Data Mining (Springer), 2016.

IndiGov-O: An ontology of Indian Government to empower Digital Governance

Akshi Kumar

Department of Computer Science and Engineering

Delhi Technological University

Delhi, India

akshi.kumar@gmail.com

Abhilasha Sharma

Department of Computer Science and Engineering

Delhi Technological University

Delhi, India.

abhilasha_sharma@yahoo.com

Arunima Joshi

Department of Computer Science and Engineering

Delhi Technological University

Delhi, India

j.arunima@yahoo.com

***Abstract*—Ontologies enable both humans and machines to communicate precisely to support the exchange of semantics by defining shared collective domain knowledge. In this paper we build an ontology using Protégé to organize knowledge of Indian Government portal. IndiGov-O, the ontology of Indian Government is a 4-level hierarchy in which various ministries of India, their respective departments and further functions and schemes have been conceptualized. The idea is to envision "digital governance" by virtue of social web adoption to a government model based on knowledge.**

Keywords— Semantic Web, Intelligent Web, Ontology, Protégé, DL(Description Logic) .

1. INTRODUCTION

Information overload and retrieval quality are two primary concerns when dealing with the current socially-hyper generation of web. Traditional search engines are unable to provide satisfactory solutions to these and thus foster the need to find, develop and implement a semantically richer web. According to Tim Berners-Lee:"The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help" [1].

The World Wide Web as a global information medium has evolved radically. The first generation, Web 1.0 was the "read only web" or "static web" where there was no user interaction or content contribution. Web 2.0 which is the "Read and Write Web" or "dynamic web" went on to establish itself as the second generation of Web. It focuses on participation and collaboration of information amongst users. Web 3.0 is the next notable change that centres on computer-to-computer interaction over the Internet. It is the intelligent

generation of web referred to as Semantic Web, the "Read Write and Execute Web" [2]. More specifically, Semantic Web is defined as a web of data which provides a common framework for data to be shared and reused across various applications. It is designed to enable reasoning and inferencing capabilities which can be added to the details of entities.

To support the exchange of information and knowledge in this extended web, ontologies are used. These are vocabularies that define the concepts and relationships with each other [3]. Ontologies are building blocks of the Semantic Web that acquire domain knowledge in a general way and result into a common understanding of a domain. These are basically specification of a conceptualization and can be shared globally. Ontologies are not dependent on the applications which use them which leads to easier maintenance of data and application.Therefore, it is a model of representing data in a given domain in organized way. Different tools are accessible & used for development of ontology, for example, Semantic-Works 2008, Swoop, OntoEdit, WebODE, Protégé etc.

In this paper, we build ontology of Indian government that conceptualizes its structure as a 4-level hierarchy in which various ministries of the government, their respective departments, departmental functions and schemes are shown [3]. The motivation to build one was clearly based on the "Digital India initiative" recently proposed & projected by the Government of India. The initiative is a big step to transform the country into a digitally empowered knowledge economy & includes projects that aim to ensure that government services