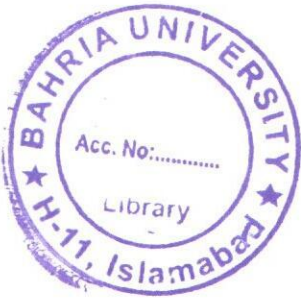


# SEMANTIC WEB MINING BY USING CLUSTERING AND PATTERN DISCOVERY



Qaneta Ahmed

01-241191-015

*Supervisor:* Dr. Tamim Ahmed Khan

A thesis submitted to the Department of Software Engineering, Faculty of Engineering Sciences, Bahria University, Islamabad in the partial fulfillment for the requirements of a Master's degree in Software Engineering

July 2021

# Approval Sheet


## Thesis Completion Certificate

Scholar's Name: Qanetah Ahmed  
Programme of MS (SE)

Registration No: 01-241191-015

Thesis Title: Semantic Web Mining By Using Clustering and Pattern Discovery

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for Evaluation. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index at 9% that is within the permissible limit set by the HEC for the MS/MPhil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/MPhil thesis.

Principal Supervisor's Signature: 

Date: 6-9-21 Name: Dr Tammam A Khan

## Certificate of Originality

This is certify that the intellectual contents of the thesis

Semantic web mining by clustering and pattern  
discovery

are the product of my own research work except, as cited property and accurately in the acknowledgements and references, the material taken from such sources as research journals, books, internet, etc. solely to support, elaborate, compare and extend the earlier work. Further, this work has not been submitted by me previously for any degree, nor it shall be submitted by me in the future for obtaining any degree from this University, or any other university or institution. The incorrectness of this information, if proved at any stage, shall authorities the University to cancel my degree.

Signature: *Qaneta*

Date: 6<sup>th</sup> Sep 2021

Name of the Research Student: Qaneta Ahmed



## Abstract

*The World Wide Web consists of different types of data present on websites and is in different formats. Different types of data include structured, unstructured and semi-structured data present in various formats. The aim of this research is to extract relational clusters from unstructured data based on sentiment by making use of natural language processing and semantic web technologies which include RDF format, FOAF ontology and OLIA ontology. Semantic web mining technologies help in converting data present online into machine readable form w.r.t ontological stand point or frameworks [45]. We use tweets in the unstructured form consisting of two columns such as person/account column and the tweet column. We convert data present into machine readable form by using natural language processing methods. The verbs extracted from data by using NLP methods are treated as predicates and the nouns/pronouns are treated as subject and object in the finalized table of person, subject, predicate and object resulting in triples. We acquire an RDF file with respective ontologies incorporated for creation of relations among triples. RDF grapher is used to visualize these relations. This study provides an in-depth analysis and implementation of how to discover meaningful patterns based on sentiment or feature the data present in unstructured form needs to be processed in-terms of machine readable form for the creation of relational clusters using ontological frameworks. The results of this study consist of ontological framework based relational data visualized in the form of clusters within clusters.*

**Keywords:** Semantic web mining, Pattern discovery, Clustering, Unstructured data, RDF graph.

# Dedication

*To my parents for their love and support*

# Acknowledgments

*In the name of Allah, the Most Gracious and the Most Merciful, all praises to Allah for the strengths and His blessing in completing this thesis. Special appreciation goes to my supervisor, Dr. Tamim Ahmed Khan, for his supervision and constant support. His invaluable help of constructive comments and guidance throughout the thesis have contributed to the success of this research.*

*My deepest gratitude goes to my beloved parents and brother; Mr Kh. Nadeem Ahmed, Mrs Saba Nadeem and M. Shaheer Ahmed for their endless love, prayers and encouragement.*

*And lastly, to the closest friend/s of mine who indirectly contributed in this research, your love, kindness and prayers means a lot to me.*

*Thank you very much.*

# Table of Contents

Approval Sheet.....	ii
Certificate of Originality .....	iii
Abstract.....	iv
Dedication.....	v
Acknowledgments .....	vi
Table of Contents .....	vii
List of Figures.....	ix
List of Tables .....	x
<b>Chapter 1 .....</b>	<b>1</b>
<b>Introduction.....</b>	<b>1</b>
1.1. Research Gap:.....	2
1.2. Problem statement: .....	2
1.3. Proposed Solution: .....	2
1.4. Research Questions: .....	3
1.5. Research objectives: .....	3
1.6. Contribution: .....	3
<b>Chapter 2 .....</b>	<b>5</b>
<b>Literature Review .....</b>	<b>5</b>
<b>Chapter 3 .....</b>	<b>12</b>
<b>Research Methodology .....</b>	<b>12</b>
3.1. Introduction .....	12
3.2. Our Methodology: .....	13
3.1.1. Data Selection and Linguistic analysis phase: .....	14
3.1.2. Triple Creation Phase: .....	14
3.1.3. Creation of relations w.r.t ontologies:.....	14
3.1.4. RDF Knowledge Graph Visualization: .....	15
<b>Chapter 4 .....</b>	<b>16</b>
<b>Implementation .....</b>	<b>16</b>
This chapter describes the steps involved in the data selection and pre-process phase, triple creation phase, creating relations w.r.t ontologies and the post process phase .....	16
4.1. Data Selection and Linguistic analysis phase.....	16
4.1.1. Data Selection: .....	16
4.1.2. Linguistic analysis phase: .....	16
4.2. Triple Creation Phase .....	17
4.2.1. Identifying Verbs: .....	17

4.2.2. Comparison, Lemmatization, Verb Net and Predicate Identification: .....	17
1. Comparison: .....	17
2. Lemmatization: .....	18
3. Verb Net and Predicate Identification: .....	18
<b>4. Subject and Object Identification: .....</b>	<b>19</b>
4.3. Creating Relations w.r.t Ontologies (FOAF and OLIA) .....	19
4.4. RDF Knowledge Graph Visualization.....	21
<b>Chapter 5 .....</b>	<b>21</b>
<b>Results and Discussion.....</b>	<b>21</b>
5.1. CSV Triples.....	22
5.2. RDF Knowledge Graph (Clusters within clusters).....	22
<b>Chapter 6 .....</b>	<b>28</b>
<b>Conclusion: .....</b>	<b>28</b>
<b>References.....</b>	<b>30</b>



# List of Figures

<b>Figure 1:</b> The Research Process.....	13
<b>Figure 2:</b> Tokenization and POS-tagging .....	16
<b>Figure 3:</b> File Read .....	17
<b>Figure 4:</b> Drafted Sentiments File.....	17
<b>Figure 5:</b> Resultant File.....	18
<b>Figure 6:</b> Resultant File (CSV) .....	19
<b>Figure 7:</b> LODRefine (Creating RDF).....	20
<b>Figure 8:</b> Person and Triples (SPO).....	22
<b>Figure 9:</b> Result set in LODRefine .....	23
<b>Figure 10:</b> Final Resultant Figure (Clusters within clusters).....	25

# List of Tables

**Table 1: FOAF and OLIA Properties.....24**

## Chapter 1

# Introduction

Data available online can be categorized as semi structured, structured and unstructured [1]. A good percentage of this data on the World Wide Web (WWW) is not structured properly. Structured data is present in a pre-defined format. It is ready to be processed and analysed by making use of data processing techniques and algorithms [44]. Whereas unstructured data (present in different formats) might be understandable by humans but for computation, analysis or prediction purposes such kind of data might not be understandable by machines. Unstructured data is difficult to process [44]. For this emerging problem making use of semantic web technologies is considered essential. Semantic web is a collection of different types of ontologies. These ontologies consist of some vocabularies that are used to identify and organize data into a meaningful form[2]. For example various user profiles from various websites are structured differently, so to make things easy the gained data of users is organized according to a specific vocabulary or syntax (the vocabulary/syntax is machine readable as well) to make it more refined and understandable for the machine and the user. Some of the main standards/frameworks that available in-terms of the semantic web technology are as follows:

- Resource Description Framework (RDF)
- Web Ontology Language (OWL)
- SPAQRL is the RDF based query language.
- Extensible Markup Language (XML)

We can use semantic web technologies to extract and identify relevant information within mixed formatted textual data sets. Data can consist of documents, user profiles or random information taken from different sources[3, 5]. The motivation behind this study is that research is on the rise in regards to using semantic web technologies [46] and NLP to find hidden patterns and relevant information based on a specific feature in unstructured data [42]. This is important to do because there is a lot of information online and it becomes difficult to keep track of information that might be related to one another based on some feature/s.

Overall most of the unstructured data requires making sense of it. In order to do so, other than applying data mining or web mining techniques, a new approach is thought of to extract relevant patterns or information based on sentiment from unstructured data/tweets and creating relations in RDF and RDF graph w.r.t ontological standpoint.

### **1.1.Research Gap:**

Researchers have analyzed and clustered data w.r.t sentiment analysis and clustering using data mining techniques [6, 9]. However it is identified that the data used for analysis in-regards to machine learning algorithms is human readable and not present in a relational format against its specific domain/feature/sentiment and remains unstructured even after results are gained [42]. For fulfilling this gap the approach of making use of NLP and semantic web technologies (FOAF and OLIA) to extract clusters of relational data based on a feature or sentiment is proposed.

### **1.2.Problem statement:**

Researchers who have worked in this field have done analysis by making use of data in unstructured form and applied machine learning algorithms including classification, prediction and clustering for getting accurate results in-terms of finding relevant data [6, 7, 33, 41]. The data used for analysis in-regards to machine learning algorithms is human readable and not present in a relational format against its specific domain/feature/sentiment and remains unstructured even after results are gained [42]. The problem in this context is that to gain meaningful information and patterns based on sentiment or feature the data present in unstructured form needs to be processed in-terms of machine readable form to gain relational clusters using ontological frameworks.

### **1.3.Proposed Solution:**

We use unstructured data to find the meaning of the tweet based on some sentiment as well as creating relations between the extracted relevant data with the account/person whom the tweet belongs to using natural language processing and semantic web technologies w.r.t ontological vocabulary and framework. However, we intend to find out if a concurrent use would be more beneficial or not.

#### **1.4.Research Questions:**

- Q1. How can we make use of Natural Language Processing to extract relevant data from unstructured form based on sentiment for creating triples?
- Q2. How can we make use of semantic web technologies to create and visualize patterns between triples w.r.t ontological standpoint?

#### **1.5.Research objectives:**

The main objectives of our research are to understand the working of semantic web ontologies and frameworks and to gather data in unstructured format, preferably in a CSV format and to extract relations based on sentiment from it in the form of clusters by making use of Natural language processing and Semantic web technologies and languages.

1. We make use of unstructured data (tweets, person/account) and by breaking down the tweet itself using stop word removal and tokenization. Then by applying part of speech tagging (based on types of verbs and nouns/pronouns), we separate the meaningful words needed for making RDF triples hence subject predicate and object, along with a column of related person to make the relation later on.
2. We create relations by making use of software LOD Refine and we use ontologies such as FOAF and OLIA to assign tags and namespaces to the gathered triples in CSV format to make sense of them. Through this process we gather triples in TURTLE format which is in machine readable form. Now by making use of rdf grapher we are able to visualize our relations acquired in TTL format, in the form of clusters within clusters w.r.t ontologies, triples and person/account relation.

#### **1.6.Contribution:**

The contribution of this work is to introduce an approach through which the pre-processing and extraction of relations from data present in unstructured form is done by incorporating Natural language processing and Semantic web technologies. We extract relevant data from unstructured data and invoke the linguistic analysis and triple creation phase which includes tokenization, part of speech tagging, lemmatization, verb net and verb comparison with a list of sentiments (based on comparison only valid tweets are extracted) and stop word removal to form triples

(subject, predicate and object). These triples are then loaded onto software known as LOD Refine where by making use of FOAF and combining it with OLIA ontology properties/tags and namespaces are assigned to create a machine processable RDF file (having relational data) and then visualized by using RDF grapher resulting in cluster within clusters of relational data.

## Chapter 2

# Literature Review

In the recent years the usage of semantic web technologies is on the rise [2, 3, 4, 27, 28]. The questions regarding how this technology can be used to work with various types of data present on the World Wide Web and to gain some relevant information/relations of data through it are increasing day by day. The usage of this technology improves the quality (in-terms of creating relations) and visualization of the related data gathered as well as without having to incorporate other machine learning algorithms; valid results are still gained. A lot of research has been done in-terms of taking data from the web and applying machine or deep learning algorithms to gain information [6, 9].

### 2.1. Clustering, Classification and Pattern Recognition

Authors in [6] have done sentiment analysis on twitter data by using a clustering algorithm, clustering the positive and negative sentiment data respectively. The author has clustered the tweets based on the sentiment/subjectivity and the polarity.

A study explained in [7] outlines the usage of opinion mining technique and calculation of the strength of each sentiment present in the data to cluster the relevant information by using K-means algorithm.

Authors have made use of sentiment analysis and clustering on data w.r.t part of speech tagging (used as training dataset) and clustering data based on a feature. To increase the accuracy of the proposed approach, Naïve bayes algorithm has also been used for the testing and training of the data w.r.t influence of the words extracted [8].

The authors in [9] have discussed comparisons between different machine learning and deep learning algorithms and how various other hybrid techniques can work to maximize the accuracy in-terms of sentiment analysis of twitter data.

An approach is proposed in this study for extraction of opinions from a text corpus (structured text) by using triplet (SPO) approach w.r.t part of speech tagging and then calculating how many times a word has occurred in the document via lexical models such as TF-IDF and to obtain feature vectors for each word Word2Vec has been used. And in the last step the author has used K-means algorithm to cluster the relevant data [10].

Natural language processing techniques alone cannot extract information needed for tweet classification. In this study the authors have proposed a hybrid approach that combines natural language processing techniques along with machine learning techniques to for the classification of twitter data [30].

In this study the authors have done a survey on detection of common interest's in-regards to real time data of twitter by using classification methods/techniques [32].

Similarly in this study the authors have compared different types of machine learning algorithms such as random forest, support vector machine, k-means clustering etc in-terms of sentiment analysis of twitter data [34].

The author in [38] made use of machine learning techniques and natural language processing to extract opinions from product reviews w.r.t behavioural and relational features.

In a similar research the authors have mentioned an approach involving natural language processing techniques on twitter data to extract tweets relevant to health related topics. The method which is used to fulfil this purpose is lexico-syntactic patterns [39].

In this study, semantic and syntactic analysis on unstructured tweets is done. Classification of tweets based on sentiment [40].

For the purpose of information gathering, an approach is proposed by the author in this study [41] which includes the integration of clustering algorithms such as k-means, DBSCAN and spectral with pattern mining algorithms applied on two use cases that include documents and twitter data.

## **2.2. NLP and Semantic Web Technologies**

Authors in [2] proposed an intelligent model which processes queries of users and based on similar searched data by other users, it semantically creates relations for further relevance and user query processing.

In another study, the authors in [3] proposed an approach for creating an RDF file (no ontology incorporated) for unstructured journals by removing stop words and simple subject, predicate and object extraction. With the use of a sparql query, showing results from the schema.

The authors in [4] have discussed details regarding how ontologies can be utilized in information extraction w.r.t common architectures and systems. Also the discussion



regarding tools for performance enhancement and calculations regarding performance using metrics has been done.

In another study, the authors proposed a technique which included named entity recognition as well as using KIMO (KIM ontology) for information extraction from knowledge base consisting of entities of general importance (person, location organization etc.) [5]. This approach validates that ontologies can be used to create relations among different types of data.

A comparison of the types of data present online is done in this study. For example structured, unstructured and semi-structured and how different types of data require specialized storing system w.r.t different characteristics. Also an approach of how can we store these types of data in RDFs form using semantic wiki KiWi (proposed by author/s) [11].

Authors in [12] have made use of eBook webpages in unstructured form, applied concept pruning for information extraction and converted the results in generic RDF format consisting of tags such as <book>, <description> etc.

In the study of semantic patterns w.r.t sentiment analysis using twitter data, the author proposed a model called SentiCircle which extracts the contextual semantic of a mentioned word and place the word under positive, neutral, very positive, negative and very negative categories based on the word's contextual occurrence in the data [13].

The authors have used twitter posts and extracted hashtags and time stamps from the tweets using an approach called as semantic patterns (pattern dictionary of English verbs) similar to verb net [21] and based on the co-occurrence of the extracted time stamps/hashtags weighted links have been assigned and the results have been showed in the form of a bar graph which validates the number of times a hashtag has occurred in a tweet in-regards to the time stamp [14].

Author/s in [15] have created their own domain ontology for their twitter dataset and have extracted attributes from the tweets based on the properties present in their own created ontology and categorized the extracted attributes w.r.t sentiment grading (positive, negative and neutral) [15].

In another study, the authors have discussed the importance of using ontologies for artificially intelligent recommendation systems in-terms of personalization and reusability [16].

Similarly in another study, the author proposed an ontological based recommendation system w.r.t machine learning techniques in the field of education [17].

The authors in this research paper have discussed various types of semantic web technologies and how these technologies can be used in other domains and potentially play an important role in doing so [18].

Authors in [19] have discussed the role of rdf data management and rdf graphs w.r.t big rdf data clouds [19].

In another study, it is discussed that how rdf data or various rdf knowledge graphs can be searched by using triple pattern queries [20] also known as sparql queries which are sample queries consisting of keywords such as select, distinct etc. Sparql queries can be used on random rdf data files to search for valid URI or entities. Searching is done by using triples (SPO) and in the where clause conditions are created with respect to the rdf data present in the file or graph [24].

Authors in [22] have done the analysis of FOAF documents in regards to social networking websites. FOAF is a vocabulary that supports friend of a friend relationship between entities present on the web. Social networks have made use foaf ontology for creation of relations of people who might follow each other or have be friended each other on social platforms. Foaf is a semantic web technology that is processable on machine level.

OLiA ontology also known as Ontologies of linguistics annotations is used to represent linguistic annotations in data. Linguistics can be linked together using OLiA ontology w.r.t lexical semantic resources/properties embedded in the ontology [23].

In this study, an ontology model based on a specific domain is proposed which focuses on document processing and retrieval of valid documents [25].

Authors in [26] have proposed their own ontology which caters to dispersing rdf triples extracted w.r.t relations based on semantic features/constraints from research publications (articles gathered from DBpedia) [26].

The authors in this study have proposed an approach of merging various user profiles on social media platforms by using the FOAF ontology semantics as well as reasoning techniques of the semantic web. Most of the social media platforms are making use of FOAF ontology to represent person to person relations in machine readable form. This creates a network of related profiles of people [27]. The authors have presented, in [27] a case study regarding FOAF ontology and how it is used to solve real life problems [28].

Since FOAF ontology can be used to create relations between people based on similar interests as well as people “knowing” one another, a network of connected people can be extracted using this ontology. Extraction of health related tweets based on three topics mentioned is done by using Natural language processing techniques and by using a lexical pattern approach proposed by the authors [29].

Natural language processing techniques alone cannot extract information needed for tweet classification. In this study the authors have proposed a hybrid approach that combines natural language processing techniques along with machine learning techniques to for the classification of twitter data [30].

In order to extract relations from social network data (twitter), the author in this paper proposed an approach which makes use of Natural language processing to create triples in CSV format and then apply machine learning techniques such as random forest and SVM to find the percentage of a relation based on occurrence of a relation in the extracted triples [31].

Authors in [33] have proposed an approach which includes extraction of twitter data in regards to sentiment polarity (topics referred: donation, charity etc.) and natural language processing.

The author made use of natural language processing framework for the filtration of tweets and by using bag of words and TF-IDF (term frequency-inverse document frequency) the analyzation of sentiment of tweets has been done [35].

The data present on twitter in the form of tweets is usually present in unstructured form due to the character limit as well as various URLs or emoticons that are included within the text. Authors in [36] have made use of natural language processing to extract keywords from tweets by using statistical methods as well as WordNet [36].

The authors proposed an approach which includes conversion of textual documents into RDF form using semantic orientation (triples) and based on topic modelling creating clusters of documents. For this purpose no ontology is incorporated to create relations among data only basic use of RDF and OWL framework [42].

Authors in [43] have proposed a model called tweeki which links twitter entities to wikidata for analysis purposes.

The authors in this study have put forth an approach for evaluation of knowledge in-terms of mega –projects carried out in the industry. For this purpose an ontology called Project Ontology that provides the base for analytics of a project is proposed in the study. [48]

Authors in [49] have made use of semantic web technologies including mining of data to propose a technique that maximizes the classification of data as-well as information search and retrieval.

The authors in [50] have done a review on the possible techniques that can be utilized for accurate information retrieval and maximization of searching accuracy including semantic technologies approach with respect to mining data by using natural language processing.

The knowledge gained through this study analysis of various types of related work, the gap of using existing ontological frameworks and NLP in-regards to posts present on social media and creating relations between them based on a feature or sentiment is clearly visible. Also the extraction of valid information from unstructured data online needs to be catered as-well. For this purpose, we have proposed an approach which takes twitter data of people (account and tweets in unstructured form), breaks the tweet down into triples using various aspects of NLP, where in SPO the predicate is the extracted sentiment and by using FOAF and OLIA we create relations to form clusters within clusters present in machine readable format.

Year	Author/s	Paper Title	Work/Purpose
2017	Ahuja, Shreya, and Gaurav Dubey	Clustering and sentiment analysis on Twitter data	Sentiment analysis on twitter data by using a clustering algorithm, clustering the positive and negative sentiment data respectively
2018	Wang, Yili, KyungTae Kim, ByungJun Lee, and Hee Yong Youn	Word clustering based on POS feature for efficient twitter sentiment analysis	Sentiment analysis on data w.r.t part of speech tagging (POS) and clustering data based on a feature. Naïve Bayes algorithm also used.
2018	Abd El-Jawad, Mohammed H., Rania Hodhod, and Yasser MK Omar	Sentiment analysis of social media networks using machine learning	The comparisons between different machine learning, deep learning algorithms and hybrid algorithms for maximization of accuracy w.r.t sentiment analysis.
2019	Riaz, Sumbal, Mehvish Fatima, Muhammad Kamran, and M. Wasif Nisar	Opinion mining on large scale data using sentiment analysis and k-means clustering	Usage of opinion mining technique and calculation of the strength of each sentiment present in the data to cluster the relevant information by using K-means algorithm.
2020	Djenouri, Youcef, Asma Belhadi, Djamel Djenouri, and Jerry Chun-Wei Lin	Cluster-based information retrieval using pattern mining	The integration of clustering algorithms such as k-means, DBSCAN and spectral with pattern mining algorithms applied data.
2020	Soukaina Fatimi, Chama El Saily and Larbi Alaoui	Semantic Oriented Text Clustering Based on RDF	The conversion of textual documents into RDF form using semantic orientation (triples) and based on topic modelling creating clusters of documents.

## Chapter 3

# Research Methodology

### 3.1. Introduction

In this chapter the research process and methodology is presented. The research process is divided into various phases which include the data selection and linguistic analysis phase, triple creation phase, creation of relations w.r.t ontologies (RDF file) and RDF file and visualization phase which includes the RDF visualization through RDF Graphs. In the first phase which is the data selection and linguistic analysis phase, selection of data is done. The data consists of two columns i.e. Person/Account and Tweet column. We take the tweets of various people from the dataset and apply tokenization and part of speech tagging (based on types of verbs and nouns/pronouns for triple purposes). In the triple creation phase we have made use of a list of various types of sentiments/emotions (happy, sad, calm, hate etc.) and have compared that list with the verbs we acquired after tokenization and part of speech tagging. (Not all verbs found in the tweets represented solid sentiments). Through this step we acquired the valid verbs or sentiments belonging to valid tweets of people. We use the valid verb (this will be considered as a predicate) and split the associated tweets into two parts. The sentence before the verb occurs, subject/s is extracted from it and similarly object/s is extracted from the part of the sentence or tweet which is present after the occurrence of the valid verb/sentiment. At the end of this phase we have triples (subject, predicate, and object) of the tweet and the person whom the tweet belongs to. In the third phase of creation of relations w.r.t ontologies, the acquired triples and person column, this data is loaded onto software known as LOD Refine. We create relations among the triples of tweets and persons by making use of FOAF (Friend of a friend) ontology and OLIA ontology (based on linguistics). Next, we create URI columns (uniform resource identifier) for the predicate column as well as the person column. Namespaces are added alongside the predicates as well as persons. FOAF ontology is incorporated with the person and OLIA ontology with the predicate column. Next, we use the tags/properties provided by the ontologies to make relations between the data. We also add relevant prefixes in the file which serve as the headers.

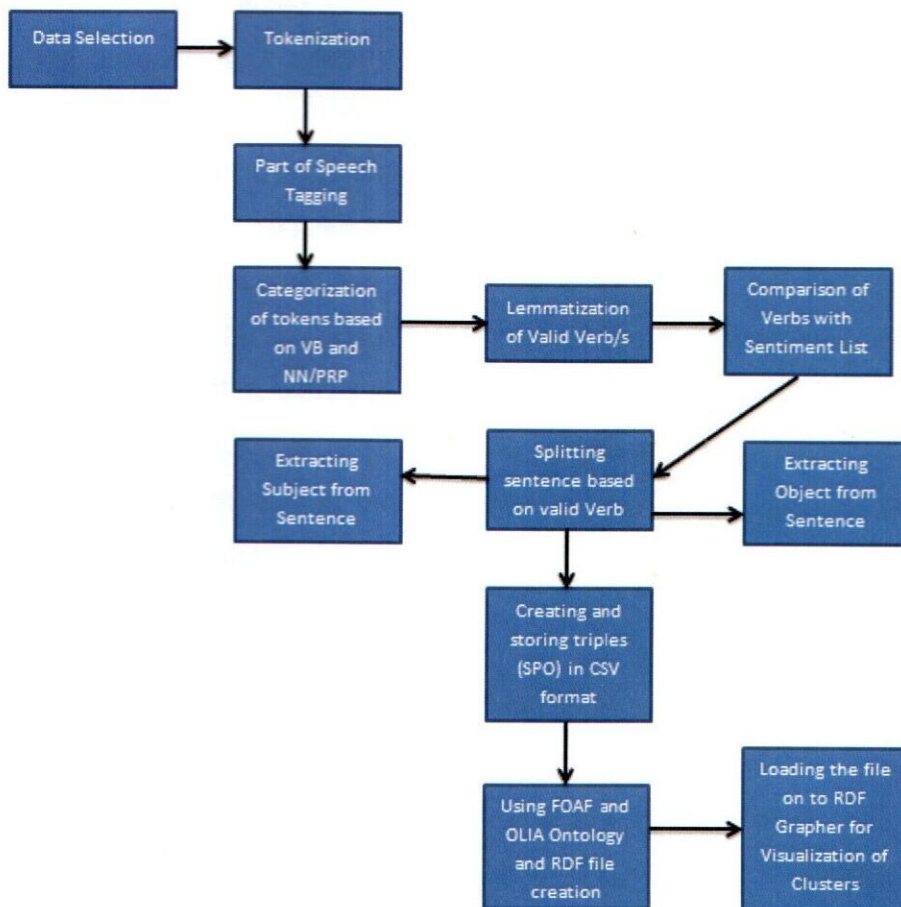
Lastly we use the TURTLE file acquired from this process and visualize it by making use of rdf grapher services.

The research approach used to conduct this research is Applied Research. We are aware of the problems that are faced when extracting relevant relations among unstructured data present on the World Wide Web. To resolve the problem, we have proposed an approach that takes unstructured data in the form of tweets and results in creating clusters within clusters based on various sentiments by making use of semantic web technologies. The various steps involved in the research process are explained below.

### 3.2. Our Methodology:

Our research process is divided in to four phases which include the pre-processing phase, triple creation phase, creation of relations w.r.t ontologies (RDF file) and post process phase which includes the RDF visualization through RDF Graphs.

In the Figure 1 given below, the steps included in our research process are shown.



**Figure 1:** The Research Process

### **3.1.1. Data Selection and Linguistic analysis phase:**

In phase one which is the data selection and linguistic analysis phase, selection of data is done. The data consists of two columns i.e. Person and Tweet column. We have taken the tweets of various people and the people whom the tweets belong to. The unlabelled dataset of tweets is publicly available on a platform in CSV format known as Harvard Data verse [47] (the dataset consisted of some extra columns which are not needed, only person and tweets column is required). A vast list of sentiments (positive, negative and neutral) is also loaded for comparison purposes later on in the process.

In the linguistic analysis phase, by making use of natural language processing tokenization and part of speech tagging is applied (based on types of verbs and nouns/pronouns for triple purposes).

### **3.1.2. Triple Creation Phase:**

In the triple creation phase we have made use of a list of various types of sentiments (positive, negative and neutral) and have compared that list with the verbs we acquired after tokenization and part of speech tagging. (Not all verbs found in the tweets represented solid sentiments) We have used lemmatization and verb net for this purpose. Through this step we have acquired the valid verbs or sentiments belonging to valid tweets of people. We use the valid verb (this will be considered as a predicate) and split the associated tweets in to two parts. The sentence before the verb occurs, subject/s will be extracted from it and similarly object/s will be extracted from the part of the sentence or tweet which is present after the occurrence of the valid verb/sentiment. At the end of this phase we will have triples (subject, predicate, and object) of the tweet and the person whom the tweet belongs to.

### **3.1.3. Creation of relations w.r.t ontologies:**

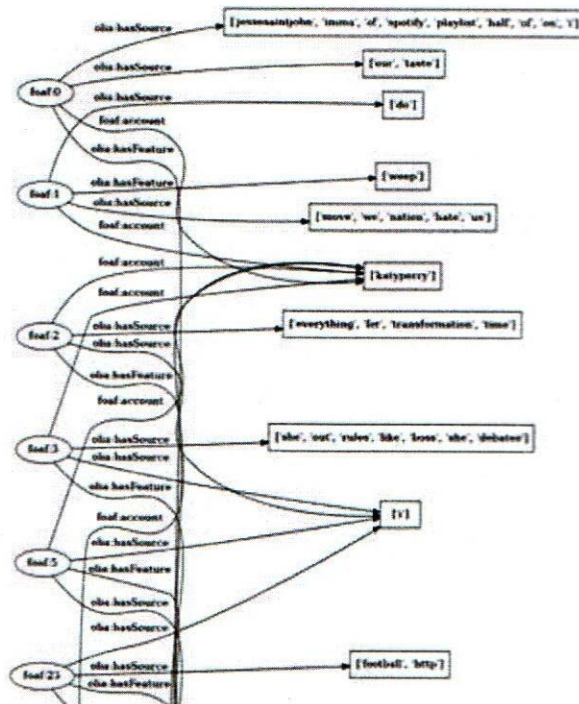
In the third phase of creation of relations w.r.t ontologies, the acquired triples and person column, this data will be loaded onto software known as LOD Refine. We create relations among the triples of tweets and persons by making use of FOAF (Friend of a friend) ontology and OLIA ontology (based on linguistics). Next, we create URI columns (uniform resource identifier) for the predicate column as well as



the person column. Namespaces are added alongside the predicates as well as persons. FOAF ontology is incorporated with the person and OLIA ontology with the predicate column. Next, we use the tags/properties provided by the ontologies to make relations between the data. We also add relevant prefixes in the file which serve as the headers. After finishing of this phase an RDF file in Turtle (.ttl) format is created consisting of ontological relations w.r.t the data (triples and person column).

### 3.1.4. RDF Knowledge Graph Visualization:

In the last phase we use the TURTLE file (machine readable) acquired from this process and visualize it by making use of online rdf grapher services to view the created relations among people and their tweets based on the extracted sentiments. Hence, resulting in clusters within clusters of the relational data.



## Chapter 4

# Implementation

This chapter describes the steps involved in the data selection and pre-process phase, triple creation phase, creating relations w.r.t ontologies and the post process phase

### 4.1. Data Selection and Linguistic analysis phase

#### 4.1.1. Data Selection:

The data for tweets and persons has been collected from one source known as Harvard Data verse [47]. The dataset consists of a person column and a tweets column. The dataset is present in CSV format.

#### 4.1.2. Linguistic analysis phase:

The pre-process phase consists of loading the data of tweets and persons. We store the tweets in a separate list and the persons in another list, and by making use of Natural language processing library (spaCy) in python, we apply tokenization and part of speech tagging on each of the tokens extracted from each unstructured tweets/sentence.

```
for sentence in tweets:
    tokn=word_tokenize(sentence)
    for word,pos in nltk.pos_tag(tokn):
```

**Figure 2:** Tokenization and POS-tagging

We also have a sentiment list (positive, negative and neutral) present in CSV format that consists of various types of sentiments such as anger, happy, happiness, sentiments related to generic on going topics on social platforms etc. This list will later on be used to extract relevant features from the unstructured tweets so that relations in the form of clusters can be made. This sentiment list is also loaded for further use.

```

sentiments=[]
with open('sentiments.csv',encoding="utf8") as csvfile:
    csvReader = csv.reader(csvfile)
    for row in csvReader:
        sentiments.append(row[0])

```

**Figure 3: File Read**

## 4.2. Triple Creation Phase

### 4.2.1. Identifying Verbs:

The identification of the verbs is done by making use of the tokens extracted from the unstructured tweets along with the pos tags of the tokens. The tokens having POS tag of any type of verb are separated from the rest of the tokens as the verbs are considered to be the predicate/sentiment/feature.

### 4.2.2. Comparison, Lemmatization, Verb Net and Predicate Identification:

#### 1. Comparison:

The extracted tokens which have pos tag of any type of verb are compared to the list of sentiments that is available. If any of the verb/token matches the sentiment present within the list, that verb/token is then lemmatized.

	A	B	C
1	fear		
2	love		
3	disgust		
4	anger		
5	sad		
6	hate		
7	happy		
8	happiness		
9	sadness		
10	grief		
11	overjoyed		
12	surprised		
13	excited		
14	scared		
15	fearful		
16	afraid		
17	friendship		
18	kindness		
19	kind		
20	pity		
21	envy		
22	shame		
23	shameful		
24	weep		
25	weeping		

**Figure 4: Drafted Sentiments File**

## 2. Lemmatization:

Lemmatization is a process that comes under text processing as well as natural language processing. The lemmatization feature is associated with NLTK in python. Lemmatization is the process of taking words having similar meaning and linking those words to one generic word that would represent all those words as a whole. For example the word better, when applied lemmatization to it, the word better will be categorized under or presented as the word good. Similarly the words plays, playing, played will be categorized under the root word called play.

In this step the purpose of using lemmatization is that if sentiments such as happier, happiness etc. occur these words will be rooted under the word happy. Hence it makes easier to group various kinds of sentiments carrying similar meaning under one word or one category. Lemmatization is applied on each word/verb acquired after comparison with the sentiment list.

```
lemmatizerrr = WordNetLemmatizer()
lemm = lemmatizerrr.lemmatize(word,pos="v")
```

## 3. Verb Net and Predicate Identification:

Verb Net is a part of NLTK in python. Verb net consists of various types of verbs categorized under meaningful classes. In this step verb net is used to identify if whether the lemmatized word/sentiment acquired makes sense or not. To achieve this purpose we use the classids function which is associated with Verb net in python. The function returns meaningful classes (if found) against the word. If the result is empty then the verb is dropped hence does not make sense. If the result contains a meaningful class which the word is categorized under, that word is stored as the finalized extracted predicate also called as the feature/sentiment on the basis of which relations between the tweets of people will be created.

```
res = [sense for sense in verbnet.classids(lemm)]
```

Verb	Lemma	VerbNet		
repeating	repeat	['say-37.7-1', 'stop-55.4']		
doin	doin	[]		
county	county	[]		
holidays	holiday	['weekend-56']		
sending	send	['confine-92-1', 'send-11.1-1']		

Figure 5: Resultant File

#### 4. Subject and Object Identification:

In this step, we use the valid verb (this will be considered as a predicate) and split the associated tweet with verb into two parts. The sentence before the verb occurs; subject/s is extracted from it based on the types of Nouns and Pronouns w.r.t the POS tagged words. Similarly object/s is extracted from the part of the sentence or tweet which is present after the occurrence of the valid verb/sentiment based on the types of Nouns and Pronouns w.r.t the POS tagged words. At the end of this phase we will have triples (subject, predicate, and object) of the tweet and the person whom the tweet belongs to in CSV format file.

Person	S	P	O
['katype['jessesair	['love']		['our', 'taste']
['katype['do]	['weep']		['move', 'we', 'nation', 'hate', 'us']
['katype['i]	['love']		['everything', 'for', 'transformation', 'time']
['katype['i]	['love']		['she', 'out', 'rules', 'like', 'boss', 'she', 'debates']
['katype['maddiso	['calm']		['i', 'room']
['katype['i]	['love']		['you']
['katype['legendai	['hate']		['typos']
['katype['legendai	['love']		['you']
['katype['conditioi	['love']		[]
['katype['limalimc	['hate']		['ok']
['katype['you', 'i']	['love']		['feature', 'instagram']
['katype['katyscru:	['calm']		[]
['katype['i]	['love']		['dives']
['katype['you', 'me	['love']		['you', 'for', 'it', 'proudfeminist']
['katype['i]	['hate']		['space', 'my', 'ocd', 'on', 'i', 'https']
['katype['skywater	['calm']		[]
['katype['my', 'mo	['surprise	['me', 'today', 'by', 'against', 'her', 'sugar', 'cereal', 'restrictions', 'https']	
['katype['girl', 'i', '	['excite']		['for', 'vmas']
['Cristia ['i', 'with',	['love']		['it', 'visit', 'http', 'http']
['Cristia ['great', 'w	['love']		['http']
['Cristia ['you']	['love']		['http', 'be', 'mercurial', 'http']
['Cristia ['you']	['hate']		['http', 'be', 'mercurial', 'http']
['Cristia ['if', 'you']	['love']		['football', 'stay', 'year', 'i', 'world', 'championship', 'you', 'your', 'friends', 'http']
['Cristia ['i']	['love']		['football', 'http']

Figure 6: Resultant File (CSV)

#### 4.3. Creating Relations w.r.t Ontologies (FOAF and OLIA)

In this phase of creation of relations w.r.t ontologies, the acquired triples and person column, this data will be loaded onto software known as LOD Refine. LOD Refine is software that is used to help in creating files in various formats. The conversion of file that is required is CSV triples to RDF triples, in-order to apply ontological properties to create relations among the data.

RDF stands for resource data framework. RDF itself carries properties which allow the merging to data present in various formats or schemas. RDF files are created by

making use of Uniform Resource Identifiers also known as URIs. The URI allows in creating a structure which links the entity to the web. Before linking the URI to the entity, the link is considered to be a namespace. This namespace when added against the entity or word creates a link of that entity with the web hence becomes a URI. After creation of the URIs, by making use of the associated properties present against the web link a linked and label file can be created. This file then can be used for visualization purposes (using rdf grapher or virtuoso server) or even SPARQL queries can be applied to it as well to extract some data from an RDF schema. The file created is an RDF file which is in TURTLE format.

AB	Person	P_URI	S	P	URI	O
10.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[malimonia, 'love', 'we']	[hate]	http://purl.org/olia/system.owl#Feature[hate]	[ok]
11.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[you, 'I']	[love]	http://purl.org/olia/system.owl#Feature[love]	[feature, 'instagar']
12.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[katyscrush, 'typosqween']	[calm]	http://purl.org/olia/system.owl#Feature[calm]	[]
13.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[I]	[love]	http://purl.org/olia/system.owl#Feature[love]	[dives]
14.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[you, 'me', 'definition', 'of', 'if', 'feminist', 'amp', 'I']	[love]	http://purl.org/olia/system.owl#Feature[love]	[you, 'for', 'if', 'proudfeminist']
15.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[I]	[hate]	http://purl.org/olia/system.owl#Feature[hate]	[space, 'my', 'tood', 'on', 'I', 'hipe']
16.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[skyywater]	[calm]	http://purl.org/olia/system.owl#Feature[calm]	[]
17.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[my, 'mother']	[surprise]	http://purl.org/olia/system.owl#Feature[surprise]	[me, 'today', 'by', 'against', 'her', 'sugar', 'cereal', 'restrictions', 'https']
18.	[katyperry]	http://xmins.com/foaf/spec/#term_account[katyperry]	[girl, 'I', 'you', 'if', 'mdmoinan']	[excite]	http://purl.org/olia/system.owl#Feature[excite]	[for, 'vmas']
19.	[Cristiano]	http://xmins.com/foaf/spec/#term_account[Cristiano]	[I, 'with', 'my', 'line']	[love]	http://purl.org/olia/system.owl#Feature[love]	[if, 'visit', 'http', 'hito']

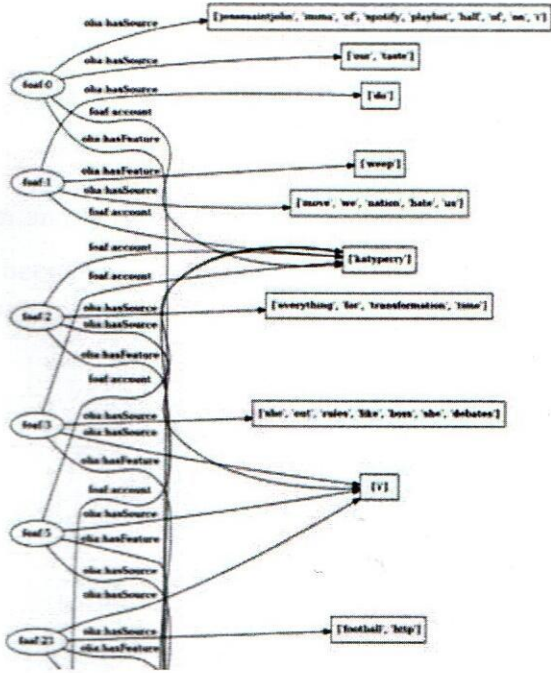
Figure 7: LODRefine (Creating RDF)

We create relations among the triples of tweets and persons by making use of FOAF (Friend of a friend) ontology and OLIA ontology (based on linguistics).

Next, we create URI columns (uniform resource identifier) for the predicate column as well as the person column. Namespaces are added alongside the predicates as well as persons. FOAF ontology is incorporated with the person/account and OLIA ontology with the predicate column. Next, we use the tags/properties provided by the ontologies to make relations between the data. We also add relevant prefixes in the file which serve as the headers. After finishing of this phase an RDF file in Turtle (.ttl) format is created consisting of ontological relations of the data (triples and person column).

#### 4.4. RDF Knowledge Graph Visualization

The relations among the data is created in the TURTLE file (rdf file). In-order to visualize the relations that are created in the machine readable file based on sentiments among people and their tweets, an online rdf graph visualization service is used known as RDF grapher. The data from the file is placed in the text box and the rdf grapher converts the file into an graphical representation in which after the incorporation of ontological frameworks clusters within clusters have been created.



### Chapter 5

## Results and Discussion

To demonstrate the validity of our proposed approach, we took the data from Harvard Data verse which consisted of people and their tweets (about 50,000+ tweets). We make sense of the unstructured tweets by splitting them into triples (subject, predicate and object) where predicate serves as the feature/sentiment which helps in creating clusters between related tweets and people whom the tweets belong to. To apply the ontological frameworks/vocabularies, software called LOD Refine was used for creation of links between the extracted data. The links are created using

properties/tags, prefixes, URI/s, namespaces etc. As per our proposed approach we are interested to extract relevant information from the unstructured tweets of people based on sentiment and creating clusters within clusters of the relevant data by using ontological vocabularies of FOAF and OLIA. The results acquired after completion of each major phase have been discussed. The first result set (CSV triples) is gained after the completion of linguistic analysis and triple creation phase. Similarly the second and final result (RDF Knowledge Graph) is extracted by making use of the first result set and after the completion of relation creation with respect to ontological phase.

### 5.1. CSV Triples

We have used tweets (unstructured) and people whom the tweets belong to, in-order to extract triples (CSV format) by making use of sentiment list (present in CSV format) and Natural language processing (tokenization, pos tagging, stop word removal, lemmatization and verb net) in python. In the image given below are some of the triples that have been extracted.

```
{'Person': ['katyperry'], 'S': ['jessesaintjohn', 'imma', 'of', 'spotify', 'playlist', 'half', 'of', 'on', 'i'], 'P': ['love', 'e'], 'O': ['our', 'taste']}
{'Person': ['katyperry'], 'S': ['do'], 'P': ['weep'], 'O': ['move', 'we', 'nation', 'hate', 'us']}
{'Person': ['katyperry'], 'S': ['i'], 'P': ['love'], 'O': ['everything', 'for', 'transformation', 'time']}
{'Person': ['katyperry'], 'S': ['i'], 'P': ['love'], 'O': ['she', 'out', 'rules', 'like', 'boss', 'she', 'debates']}
{'Person': ['katyperry'], 'S': ['maddisonxperry', 'women'], 'P': ['calm'], 'O': ['i', 'room']}
{'Person': ['katyperry'], 'S': [], 'P': ['calm'], 'O': []}
{'Person': ['katyperry'], 'S': ['i'], 'P': ['love'], 'O': ['you']}
{'Person': ['katyperry'], 'S': ['lesbiyonce', 'i'], 'P': ['love'], 'O': ['kanye']}
{'Person': ['katyperry'], 'S': ['legendarymalek', 'crap', 'i', 'i'], 'P': ['hate'], 'O': ['typos']}
{'Person': ['katyperry'], 'S': ['legendarymalek', 'bb', 'i', 'been', 'book', 'bearer', 'of', 'done', 'come', 'for', 'me'], 'P': ['love'], 'O': ['you']}
{'Person': ['katyperry'], 'S': ['conditionalbabe', 'for', 'me', 'i'], 'P': ['love'], 'O': []}
{'Person': ['katyperry'], 'S': ['limalimoncia', 'love', 'we'], 'P': ['hate'], 'O': ['ok']}
{'Person': ['katyperry'], 'S': ['you', 'i'], 'P': ['love'], 'O': ['feature', 'instagram']}
{'Person': ['katyperry'], 'S': ['katyscrush', 'typosqueen'], 'P': ['calm'], 'O': []}
{'Person': ['katyperry'], 'S': ['i'], 'P': ['love'], 'O': ['dives']}
{'Person': ['katyperry'], 'S': ['you', 'me', 'definition', 'of', 'it', 'feminist', 'amp', 'i'], 'P': ['love'], 'O': ['you', 'for', 'it', 'proudfeminist']}
```

Figure 8: Person and Triples (SPO)

### 5.2. RDF Knowledge Graph (Clusters within clusters)

To make relations between the extracted triples as shown above, we made use of FOAF and OLIA ontology where FOAF ontology represents person to person relationship, person to account relationship etc. and the OLIA ontology represents the linguistics aspect of the relation which in this scenario occurs between the triples (subject, predicate and object). Predicate is considered to be the feature on the basis of which clusters/relations are created. Subject and object represent the valid words



found in the tweets/sentences. In-order to make our approach effective, we have incorporated the ontologies by using software called LOD Refine. LOD Refine helps to load the ontology onto a platform so that it may be utilized accordingly. The following image validates the csv triples being utilized and assigned namespaces and URI columns along with properties and prefixes w.r.t FOAF and OLIA ontology.

All	Person	P_URI	S	P	URI	O
10	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[malmonica, love, we]	[hate]	http://purl.org/olia/system.owl#Feature[hate]	[ok]
11	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[you, I]	[love]	http://purl.org/olia/system.owl#Feature[love]	[feature, instagar]
12	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[katycrush, typosween]	[calm]	http://purl.org/olia/system.owl#Feature[calm]	[]
13	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[I]	[love]	http://purl.org/olia/system.owl#Feature[love]	[dives]
14	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[you, me, definition, of, I, feminat, amp, I]	[love]	http://purl.org/olia/system.owl#Feature[love]	[you, for, I, proudleinst]
15	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[I]	[hate]	http://purl.org/olia/system.owl#Feature[hate]	[space, my, ood, br, I, tops]
16	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[skyywater]	[calm]	http://purl.org/olia/system.owl#Feature[calm]	[]
17	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[my, mother]	[surprise]	http://purl.org/olia/system.owl#Feature[surprise]	[me, today, by, again, her, sugar, cereal, restrictions, http]
18	[katyperry]	http://xmlns.com/foaf/spec/#term_account[katyperry]	[girl, I, you, I, mdmolan]	[excite]	http://purl.org/olia/system.owl#Feature[excite]	[for, vmas]
19	[Cristiano]	http://xmlns.com/foaf/spec/#term_account[Cristiano]	[I, with, my, line]	[love]	http://purl.org/olia/system.owl#Feature[love]	[I, wait, http, http]
20	[Cristiano]	http://xmlns.com/foaf/spec/#term_account[Cristiano]	[great, win, night]	[love]	http://purl.org/olia/system.owl#Feature[love]	[http]
21	[Cristiano]	http://xmlns.com/foaf/spec/#term_account[Cristiano]	[you]	[love]	http://purl.org/olia/system.owl#Feature[love]	[http, be, mercurial, http]
22	[Cristiano]	http://xmlns.com/foaf/spec/#term_account[Cristiano]	[you]	[hate]	http://purl.org/olia/system.owl#Feature[hate]	[http, be, mercurial, http]
23	[Cristiano]	http://xmlns.com/foaf/spec/#term_account[Cristiano]	[I, you]	[love]	http://purl.org/olia/system.owl#Feature[love]	[football, stay, year, I, world, championshp, you, your, friend, http]
24	[Cristiano]	http://xmlns.com/foaf/spec/#term_account[Cristiano]	[I]	[love]	http://purl.org/olia/system.owl#Feature[love]	[football, http]

Figure 9: Result set in LODRefine

After the assignment of properties, prefixes etc. the RDF file consisting of triples is created as shown below:

Property/Relation	Usage
foaf:account	This property is used to represent the person or account (of the person) to whom the tweet belongs to.
foaf:id/number	The id or number is auto generated. It assigns a random id to the account of the person or person

	themselves. For example foaf:1, foaf:2
olia:hasSource	In this scenario, the hasSource property is being utilized for creating a relation between the person and the actual targeted source which is the tweet itself of the person/account.
olia:hasFeature	The hasFeature property is utilized w.r.t the hasSource property. From within the targeted source, a relevant feature (targeted object) is extracted to create a relation between them.

**Table 1:** FOAF and OLIA Properties

In-order to validate and verify if the relations among triples and persons have been created and are correct, we have used the RDF grapher service to view the created relations between the extracted relevant data by uploading our TURTLE (rdf) file. The rdf grapher (online service) [38] converts the file into a visualization form as shown in the image below:

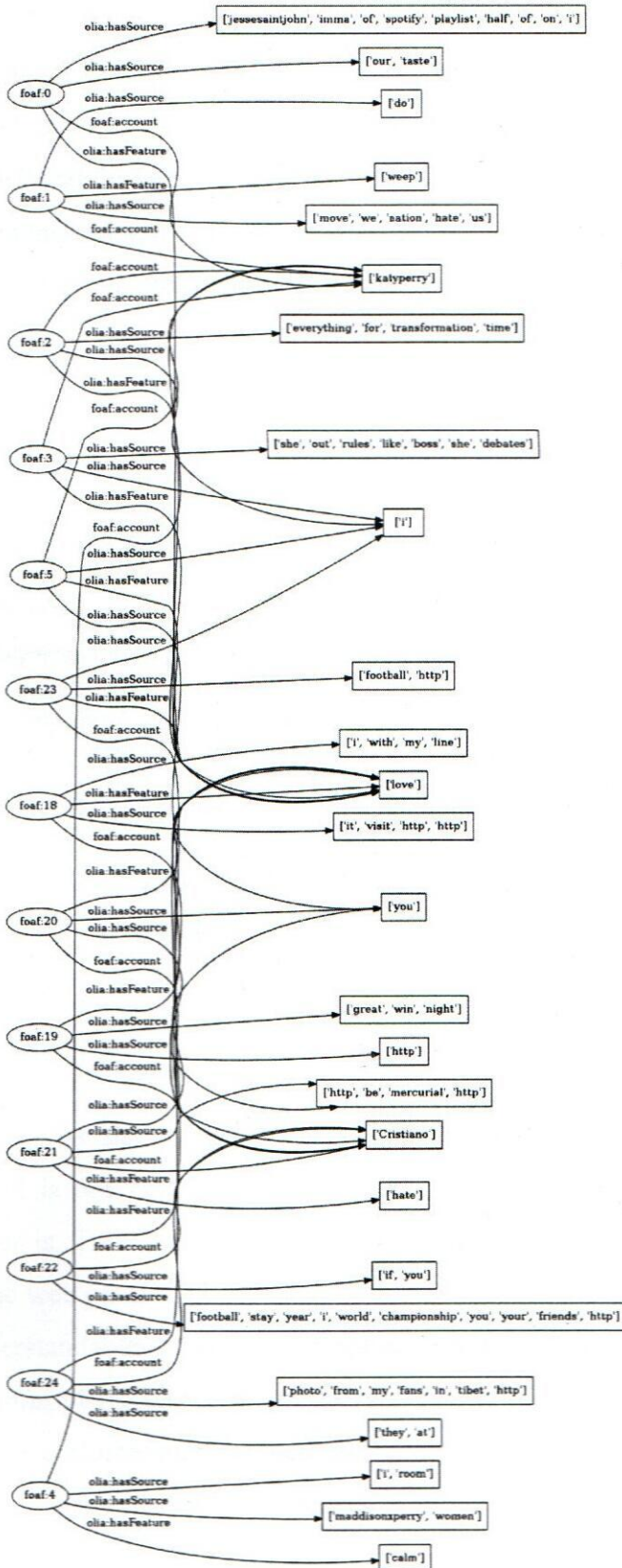
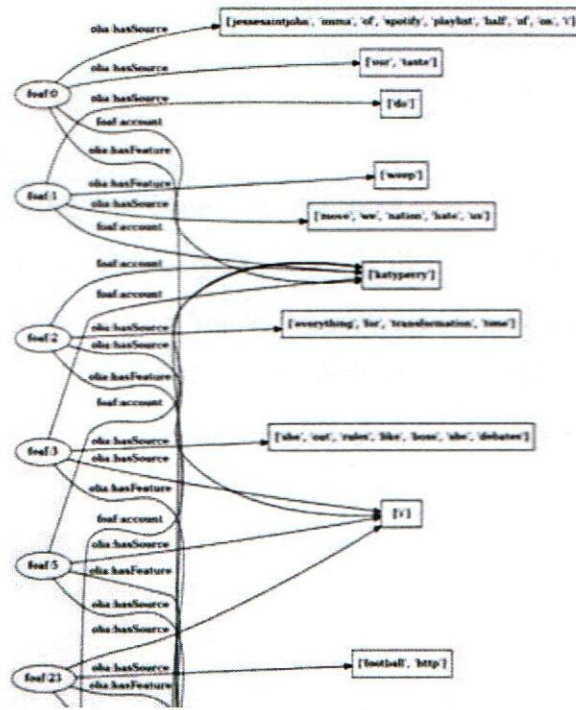


Figure 10: Final Resultant Figure (Clusters within Clusters)



In another comparative analysis, the author/s in this study [42] have made use of ontological technologies to do text clustering of web documents. The work aims at clustering of certain text based phrases which serve as a base feature for the clustering of the related documents. Ontology based languages such as OWL and RDF have been used to implement the concept of clustering of documents using text phrases. Although the concept of clustering is implemented using OWL but the relation between entities of whom this document belongs to, which other person has a similar document present online consisting of similar phrases; the relations between data through which a certain pattern can be found or implemented is not done. In our work we have made use of FOAF and OLIA ontology to implement the concept of clustering as-well as pattern recognition. These technologies have been applied to twitter data [47] which consists of a person column and a tweet/content column (only the required columns are used). The gained results from our work show the creation of relations among related entities based on extracted sentiment. For example the person whom the tweets belongs to, the sentiment hidden in the tweet, if there is any other person whom has a tweets with the same sentiment; all these links are created in the resultant figure as shown above.

## Chapter 6

### Conclusion:

This chapter discusses the conclusion of our work on semantic web technologies: clustering and pattern recognition by making use of natural language processing as well as semantic web technologies including ontological languages/frameworks, the results and overall summarization of our research work. In addition, this chapter also provides an overview of performed research and future work. We proposed an efficient approach for extracting relevant relations from unstructured and unlabelled data by making use of natural language processing and semantic web technologies and to make the extracted relations be in machine readable format.

The motivation behind this proposed approach is that data remains unstructured, not relational and not machine readable when used in-regards to machine learning [42]. NLP and semantic web technologies are used to find hidden patterns and relevant information based on a specific sentiment in unstructured data. There is a lot of information online and it becomes difficult to keep track of information that might be related to one another based on some feature.

Subject, predicate and object are acquired by splitting up the tweets in the dataset and extracting only useful and needed words/sentiments to create relations along with the person whom the tweet belongs to. When our approach is compared to K-means clustering or clustering algorithms of machine learning, such algorithms are efficient enough to group entities based on a certain feature. The lacks in that approach are; we are unable to group and create proper relations between the grouped entities on a machine readable level. For example in K-means algorithm, clusters are formed but the relations among those clusters based on a sentiment or opinion or a feature cannot be formed or are visible in the form of clusters within clusters [6, 9, 10]. Another lack is that the results gained from clustering and pattern recognition algorithms in-terms of machine learning are not in machine readable format.

Our approach focuses on creating the relations among the relevant data based on sentiments, related tweets of people having similar sentiment are linked together and are present in machine readable format as well (.ttl file). The need for including a sentiment analysis algorithm was excluded because the training and testing of the data is not required as well as based on sentiment polarity, data can only be classified

under negative, positive and neutral categories. We require solid sentiments present within a sentence, on the basis of which clusters within clusters are created.

Hence to fill this gap a list of sentiments is created in CSV format having more than 70 sentiments. No specific library dedicated to a list of sentiments is found in-terms of python for example a library equivalent to verb net.

Based on our result and experiment, we observe that we can extract relevant relations from unstructured data (tweets) by incorporating semantic web technologies.

For the future work, we are planning to take other kinds of unstructured data as-well and extract relevant information based on hashtags, topics or opinions w.r.t natural language processing and semantic web ontologies and we are planning to create and apply sparql query/s for the machine readable files to extract data from various formats of RDF files.

## REFERENCES

1. Singh, Satyaveer, and Mahendra Singh Aswal. "Semantic Web Mining: Survey and Analysis." *Journal of Web Engineering & Technology* 5, no. 3 (2018): 20-31.
2. Kabir, Sumaiya, Shamim Ripon, Mamunur Rahman, and Tanjim Rahman. "Knowledge-based data mining using semantic web." *IERI Procedia* 7 (2014): 113-119.
3. Gandhi, Kalgi, and Nidhi Madia. "Information extraction from unstructured data using RDF." In *2016 International Conference on ICT in Business Industry & Government (ICTBIG)*, pp. 1-6. IEEE, 2016.
4. Wimalasuriya, Daya C., and Dejing Dou. "Ontology-based information extraction: An introduction and a survey of current approaches." (2010): 306-323.
5. Popov, Borislav, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, and Miroslav Goranov. "Towards semantic web information extraction." In *Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003)*, vol. 20. 2003.
6. Ahuja, Shreya, and Gaurav Dubey. "Clustering and sentiment analysis on Twitter data." In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, pp. 1-5. IEEE, 2017.
7. Riaz, Sumbal, Mehvish Fatima, Muhammad Kamran, and M. Wasif Nisar. "Opinion mining on large scale data using sentiment analysis and k-means clustering." *Cluster Computing* 22, no. 3 (2019): 7149-7164.
8. Wang, Yili, KyungTae Kim, ByungJun Lee, and Hee Yong Youn. "Word clustering based on POS feature for efficient twitter sentiment analysis." *Human-centric Computing and Information Sciences* 8, no. 1 (2018): 1-25.
9. Abd El-Jawad, Mohammed H., Rania Hodhod, and Yasser MK Omar. "Sentiment analysis of social media networks using machine learning." In *2018 14th international computer engineering conference (ICENCO)*, pp. 174-176. IEEE, 2018.
10. Feldman, Daniil. "Using subject-predicate-object triplets for opinion mining."
11. Sint, Rolf, Sebastian Schaffert, Stephanie Stroka, and Roland Ferstl. "Combining unstructured, fully structured and semi-structured information in semantic wikis." In *CEUR Workshop Proceedings*, vol. 464, pp. 73-87. 2009.
12. Uddin, Ashraf, Rajesh Piryani, and Vivek Kumar Singh. "Information and relation extraction for semantic annotation of ebook texts." In *Recent Advances in Intelligent Informatics*, pp. 215-226. Springer, Cham, 2014.
13. Saif, Hassan, Yulan He, Miriam Fernandez, and Harith Alani. "Semantic patterns for sentiment analysis of Twitter." In *International Semantic Web Conference*, pp. 324-340. Springer, Cham, 2014.
14. Teufl, Peter, and Stefan Kraxberger. "Extracting semantic knowledge from twitter." In *International Conference on Electronic Participation*, pp. 48-59. Springer, Berlin, Heidelberg, 2011.
15. Kontopoulos, Efstratios, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. "Ontology-based sentiment analysis of twitter posts." *Expert systems with applications* 40, no. 10 (2013): 4065-4074.
16. George, Gina, and Anisha M. Lal. "Review of ontology-based recommender systems in e-learning." *Computers & Education* 142 (2019): 103642.
17. Obeid, Charbel, Inaya Lahoud, Hicham El Khoury, and Pierre-Antoine Champin. "Ontology-based recommender system in higher education." In *Companion Proceedings of the The Web Conference 2018*, pp. 1031-1034. 2018.
18. Patel, Archana, and Sarika Jain. "Present and future of semantic web technologies: a research statement." *International Journal of Computers and Applications* (2019): 1-10.
19. Elzein, Nahla Mohammed, Mazlina Abdul Majid, Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Fadele Ayotunde Alaba, and Muhammad Imran. "Managing big RDF data in clouds: Challenges, opportunities, and solutions." *Sustainable Cities and Society* 39 (2018): 375-386.
20. Arnaout, Hiba, and Shady Elbassuoni. "Effective searching of RDF knowledge graphs." *Journal of Web Semantics* 48 (2018): 66-84.

21. Verbnets introduction and guidelines: [https://verbs.colorado.edu/verb-index/VerbNet\\_Guidelines.pdf](https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf)
22. Ding, Li, Lina Zhou, Tim Finin, and Anupam Joshi. "How the semantic web is being used: An analysis of foaf documents." In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 113c-113c. IEEE, 2005.
23. Chiarcos, Christian, and Maria Sukhareva. "Olia—ontologies of linguistic annotation." *Semantic Web* 6, no. 4 (2015): 379-386.
24. Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. "Semantics and complexity of SPARQL." *ACM Transactions on Database Systems (TODS)* 34, no. 3 (2009): 1-45.
25. Yu, Binbin. "Research on information retrieval model based on ontology." *EURASIP Journal on Wireless Communications and Networking* 2019, no. 1 (2019): 1-8.
26. Pertsas, Vayianos, and Panos Constantopoulos. "Ontology-driven information extraction from research publications." In *International Conference on Theory and Practice of Digital Libraries*, pp. 241-253. Springer, Cham, 2018.
27. Golbeck, Jennifer, and Matthew Rothstein. "Linking Social Networks on the Web with FOAF: A Semantic Web Case Study." In *AAAI*, vol. 8, pp. 1138-1143. 2008.
28. Graves, Mike, Adam Constabaris, and Dan Brickley. "Foaf: Connecting people on the semantic web." *Cataloging & classification quarterly* 43, no. 3-4 (2007): 191-202.
29. Doan, Son, Elly W. Yang, Sameer Tilak, and Manabu Torii. "Using natural language processing to extract health-related causality from twitter messages." In *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pp. 84-85. IEEE, 2018.
30. Stavrianou, Anna, Caroline Brun, Tomi Silander, and Claude Roux. "NLP-based feature extraction for automated tweet classification." *Interactions between Data Mining and Natural Language Processing* 145 (2014).
31. Adriani, Marco, Marco Brambilla, and Marco Di Giovanni. "Extraction of Relations Between Entities from Human-Generated Content on Social Networks." In *International Conference on Web Engineering*, pp. 48-60. Springer, Cham, 2019.
32. Hasan, Mahmud, Mehmet A. Orgun, and Rolf Schwitter. "A survey on real-time event detection from the twitter data stream." *Journal of Information Science* 44, no. 4 (2018): 443-463.
33. Desai, Radhi D. "Sentiment Analysis of Twitter Data." In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 114-117. IEEE, 2018.
34. Al-Hadhrami, Suheer, Norah Al-Fassam, and Hafida Benhidour. "Sentiment analysis of english tweets: A comparative study of supervised and unsupervised approaches." In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1-5. IEEE, 2019.
35. Hasan, Md Rakibul, Maisha Maliha, and M. Arifuzzaman. "Sentiment Analysis with NLP on Twitter Data." In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pp. 1-4. IEEE, 2019.
36. Jayasiriwardene, Thiruni D., and Gamage Upeksha Ganegoda. "Keyword extraction from Tweets using NLP tools for collecting relevant news." In *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 129-135. IEEE, 2020.
37. Rdf grapher service: <https://www.ldf.fi/service/rdf-grapher>
38. Tanwar, Poonam, and Priyanka Rai. "A proposed system for opinion mining using machine learning, NLP and classifiers." *IAES International Journal of Artificial Intelligence* 9, no. 4 (2020): 726.
39. Doan, Son, Elly W. Yang, Sameer S. Tilak, Peter W. Li, Daniel S. Zisook, and Manabu Torii. "Extracting health-related causality from twitter messages using natural language processing." *BMC medical informatics and decision making* 19, no. 3 (2019): 79.
40. Khattak, Asad Masood, Rabia Batool, Fahad Ahmed Satti, Jamil Hussain, Wajahat Ali Khan, Adil Mehmood Khan, and Bashir Hayat. "Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation." *Complexity* 2020 (2020).
41. Djenouri, Youcef, Asma Belhadi, Djamel Djenouri, and Jerry Chun-Wei Lin. "Cluster-based information retrieval using pattern mining." *Applied Intelligence* (2020): 1-16.



42. Fatimi, Soukaina, Chama El Saili, and Larbi Alaoui. "Semantic Oriented Text Clustering Based on RDF." In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1-8. IEEE, 2020.
43. Harandizadeh, Bahareh, and Sameer Singh. "Tweeki: Linking Named Entities on Twitter to a Knowledge Graph." In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pp. 222-231. 2020.
44. Structured and Unstructured Data: <https://learn.g2.com/structured-vs-unstructured-data>
45. Semantic web: [https://en.wikipedia.org/wiki/Semantic\\_Web](https://en.wikipedia.org/wiki/Semantic_Web)
46. Patel, Archana, and Sarika Jain. "Present and future of semantic web technologies: a research statement." *International Journal of Computers and Applications* (2019): 1-10.
47. Dataset Link:  
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JBXXFD>
48. Zangeneh, Pouya, and Brenda McCabe. "Ontology-based knowledge representation for industrial megaprojects analytics using linked data and the semantic web." *Advanced Engineering Informatics* 46 (2020): 101164.
49. Murtaza, Sana, and Sajjad Ahmed. "Impact of the Semantic Web mining by using different techniques-A Survey." (2020).
50. Paithane, Pradip M., S. N. Kakarwal, and Sushant S. Khedgikar. "Semantic Web Technology and Data Mining for Personalized System to Online E-Commerce." *International Journal of Progressive Research in Science and Engineering* 1, no. 5 (2020): 136-139.

