Enriching Domain Concepts with Qualitative Attributes: A Text Mining based Approach

Niyati Kumari Behera and Guruvayur Suryanarayanan Mahalakshmi
Department of Computer Science and Engineering, Anna University, India

Abstract: *Attributes, whether qualitative or non-qualitative are the formal description of any real-world entity and are crucial in modern knowledge representation models like ontology. Though ample evidence for the amount of research done for mining non-qualitative attributes (like part-of relation) extraction from text as well as the Web is available in the wealth of literature, on the other side limited research can be found relating to qualitative attribute (i.e., size, color, taste etc.,) mining. Herein this research article an analytical framework has been proposed to retrieve qualitative attribute values from unstructured domain text. The research objective covers two aspects of information retrieval (1) acquiring quality values from unstructured text and (2) then assigning attribute to them by comparing the Google derived meaning or context of attributes as well as quality value (adjectives). The goal has been accomplished by using a framework which integrates Vector Space Modelling (VSM) with a probabilistic Multinomial Naive Bayes (MNB) classifier. Performance Evaluation has been carried out on two data sets (1) HeiPLAS Development Data set (106 adjective-noun exemplary phrases) and (2) a text data set in Medicinal Plant Domain (MPD). System is found to perform better with probabilistic approach compared to the existing pattern-based framework in the state of art.*

Keywords: *Information retrieval, text mining, qualitative attribute, adjectives, natural language processing.*

Received July 24, 2019; accepted May 4, 2020
https://doi.org/10.34028/iajit/17/6/10

1. Introduction

Physical attributes play a vital role in describing any real-world object. They can be broadly classified into two types:

1. Non-relational attributes.
2. Relational attributes as described in [10].

Non-relational attributes mainly include the 'part-of' relations of the object. For example, 'legs', 'beak' and 'wing', are examples of such attributes for the concept 'bird'. Apart from non- relational attributes, there are certain inherent characteristics of concepts like 'color', 'texture', 'size', 'taste', 'weight', 'shape' etc., which provide a formal description of any real-worldentity. These concept descriptors are also known as relational or Qualitative Attributes (QA). Adjectives are one of the prime contributors to describe such properties. Identifying or determining attribute concept for adjectives can refine ones understanding of natural language sentence [4] as certain adjectives can exhibit more than one property of an object. For instance, adjective 'huge' can describe a noun with respect to two different property 'weight' and 'size'. Additionally, attribute identification can also help to disambiguate among various senses of a single adjective as in the examples 'hot water' and 'hot topic'.

In natural language, adjective noun pairs are the most frequently used linguistic patterns to assign attributes to entities. Learning concept attribute as instigated by [2, 3] aimed to learn attribute-value pairs from adjective-noun phrases in natural language text. For example, the adjective in the phrase 'a red flower' describes the color of the flower. In another instance, the adjective in the phrase 'a big house' describe the size of the house. Similarly, adjectives like 'sweet', 'salty', 'sour' describe about the taste property associated with any object. Adjectives like 'gigantic', 'small' or 'long' etc. speak about the magnitude of an entity. Identifying such QAs for adjectives can revamp our understanding of real-world entities in a natural language sentence. On the other hand, in case of relation learning where aim is to learn non-taxonomic relation between semantic concepts, relational adjectives provide valuable information [6]. For example, adjective in the phrase 'a musical instrument', indicates that the semantic relation instrument to be used in music.

Most research work conducted in the area of structured domain model like ontology creation, focus on components such as concept, relation, axioms and instances. In the literature, several models have been discussed for concept extraction [9, 15, 24, 27, 28] and relation extraction [8, 17, 19, 20, 26]. Natural Language Processing (NLP) phrases like verbs and noun phrases have been widely discussed in these cases. However, though adjectives have been less discussed in this context, [1] have used relational adjectives to extract hyponyms from medical domain text. Adjectives have been analysed to learn concept attributes for ontology induction by Almuhareb and

Poesio [2] Cimiano [7]. Adjective and noun pairs as illustrated in the above examples, are prime source for both type of learning task. However, it is essential to distinguish between property denoting and relational adjectives for both relation and attribute learning.

In this paper, attempts have been made to come up with a model to ascribe attribute to adjectives in general. The basic approach followed in this work is to first identify these basic properties of real word entities present in the unstructured text and then classify them into appropriate attribute or quality type. As the ultimate goal of this research is to enrich a medicinal plant domain ontology with all inherent characteristics of medicinal plants, the domain under taken in this paper for analysis is Medicinal Plants. In Medicinal Plant Domain (MPD) these property denoting adjectives are referred as Organoleptic features or attributes of plants. Features like shape, size, odour, taste, texture, and color can act as a determinant in their differential uses for treating various diseases. Commonality among medicinal plants can be identified by analysing their organoleptic features.

This paper presents an NLP based framework to acquire set of attributes that an adjective can point to by analysing their meaning or context retrieved from the Web. Here, focus has been restricted to identify only property denoting adjectives as they are essential for efficient concept representation for ontology learning. The outcome of this research work can bring about semantic description of adjectives required to devise new techniques for capturing information about domain concepts.

With this brief introduction about the objective of this paper, organization of subsequent parts is follows: section 2 reviews the existing works related to adjective classification whereas section 3 elaborates on the quality value extraction and classification methodologies used in the proposed framework. Finally, detailed performance analysis has been discussed in section 4 followed by conclusion in section 5.

2. Related Work

Linguistically, QAs are the adjectives used to describe a common noun. In the wealth of literature, several models have been proposed to extract these attributes and their values from structured, semi structured as well as unstructured text. In the state of arts, adjective-noun phrases are found to be the most frequently and widely used linguistic pattern to impute those relational attributes. Most of the models extract knowledge about attributes using pre-defined syntactic patterns [3, 7] from large corpora like British National Corpus (BNC) and Web.

Hatzivassiloglou and McKeown [14] performed the first venture towards identifying adjective scale. They explored the linguistic information about adjectives and used them to cluster the adjectives with same scale based on their orientation. Similar linguistic information has been used in by Almuhareb and Poesio [2] Poesio and Almuhareb [23] to extract concept descriptors in terms of both attribute names as well as values. An unsupervised and domain independent framework proposed by Sánchez [25] enriches ontological concepts with attributes and their property details. Here the Web has been used as the learning corpus to extract data and to hypothesize knowledge distribution using efficient contextualized user queries. A probabilistic framework [16] with typicality score for attributes of concepts employs heterogeneous data sources like available knowledgebase, web documents and search logs, to derive the scores by aggregating their distributions in different sources. Though this novel technique is effective in identifying wide range of concept attributes, still it involves complex mathematical computation.

An architecture developed by Boleda [5] identifies semantic classes for adjectives by incorporating clustering and decision tree technique to group the adjectives under three class labels (BEO) “Basic Adjectives, Event-related Adjectives, Object-related Adjectives. An unsupervised clustering approach based on co-occurrence of adjective and noun has been proposed [22] to induce set of adjectives to the value space of an attribute. A framework by identifying the noun semantics has been proposed Hartung and Frank [12] to retrieve attributes for adjective-noun phrases. The framework stands on the class labelling method suggested by Torrent with a variation in the technique used. Authors have proposed a semi supervised classification scheme rather than clustering for the targeted acquisition. They have broadly classified the adjectives into property and relation denoting adjectives for ontology learning. Another work [11] proposed by them for attribute learning examines the doublet co-occurrences i.e., first search for noun-attribute co-occurrences and then adjective-attribute co-occurrences. However, doublet co-occurrences do not result in significant boost in web hits for patterns and their approach still lacks breadth in identifying adjective attributes. In another attempt [13] word embedding techniques have been employed to infer attribute for an adjective-noun phrase.

Though authors have registered to outperform the count-based models [11] in attribute prediction task, for computational simplicity, here analysis has been restricted to adjective approaches of attribute learning from text analysis.

A model very similar to our approach has been proposed by Bakhshandeh and Allen [4] where learn property denoting attributes by reading glosses of seed adjectives. They have initiated the algorithm with attribute details of 620 adjectives in WordNet as seed

and used bootstrapping to learn attributes for the remaining adjectives by pattern learning. It also focuses on ordering of adjective intensities. Our work fundamentally differs from this approach by the fact that instead of using glossary of seed adjectives, directly glossary of the attribute concept has been used in the proposed model.

From the literature review, it is apparent that there are still scopes to explore extraction techniques for qualitative attributes and also work can be done to reduce the computational overhead in existing models for the same. Compared to the existing works, the proposed framework is simple. The model is tailored to take into account the semantic definition of the adjectives as well as qualitative attributes directly from Wordnet derived using PyDictionary (), a Dictionary Module for Python 2/3. It then employs a Vector Space Model (VSM) based representation and probabilistic approach to compare and classify the property denoting or qualitative adjectives into their respective attribute type. This scheme overcomes the overhead added due to the manual annotation during training phase of supervised and semi supervised learning.

3. Automatic Quality Value Classifier (AQVC)

In this section, the proposed system architecture for extracting and classifying relational or qualitative attributes values associated with a concept has been elaborated. As shown in Figure 1, the knowledge acquisition process starts with a pre-processed domain corpus, a set of domain concepts and a set of predefined patterns as input. The domain corpus comprises a set of text documents acquired from various online web resources. Complete description about the corpus is given in the performance evaluation section. The basic approach followed here is to locate those qualitative adjectives associated with domain concepts from unstructured text and then employ trained classifier to ascribe attribute concept to them. The classifier analyses and compares the semantic descriptions of both adjectives and attributes as retrieved from Google before suggesting an attribute type for the adjective. Prior works have dealt with only distribution of adjective and Noun pairs using syntactic patterns. But here, in addition to the regular pattern-based information extraction, the objective is to explore their semantic descriptions to achieve more precise result.

The major components of Quality Value Detection (AQVC) are explained in brief below:

- *Quality Context Knowledgebase (QCKB)*: QCKB is built by retrieving and storing meanings of QAs from web using the above-mentioned python module. Those meanings are then pre-processed, represented in VSM and stored in QCKB. The entire knowledgebase creation process has been discussed in section 3.2.
- *Quality value extraction*: A set of pre-defined patterns named as Quality Value Detection (QVD) patterns are used to retrieve a set of Adjective-Noun (A-N) pairs from the corpus based on the sentence structure. A detail insight about the QVD patterns is given in section 3.3
- *Concept-adjective pair Extraction*: With a set of pre-defined domain concepts as nouns, the A-N pairs obtained in the previous stepare further filtered to retain the pairs having domain concepts only.
- *Web-based context extraction*: Google meanings of the adjectives present in the above filtered pairs are retrieved from web using Py Dictionary() module of python. This python module uses WordNet to get the meaning of a word. Meanings retrieved from Google describe the various context where the word can be used.
- *Context Similarity Computation*: Finally, a vector representation of meanings of both attribute and adjective are compared using trained classifier models to assign the former to the later.

Taking functionality into account, the key components of AQVC are as follows:

1. Quality Value Extractor (QVE): This component comprises of the Quality value extraction and Concept-Adjective pair extraction blocks of the architecture. It generates a set of candidate Concept-Adjective pair for classification.
2. Quality Value Classifier (QVC): It comprises of the QCKB, web context extraction and context similarity computation blocks of architecture. It classifies the candidate Concept-Adjective pairs to assign attribute type to the adjectives.

The overall architectural efficiency of the framework depends on the combined capability of QVE as well as QVC.QVE should have the potential to deal with various sentence structures in corpus whereas QVC should be able to correctly classify the quality values in to respective abstract classes. As an initial step to attain the research objective, QCKB is created as discussed in the following section.

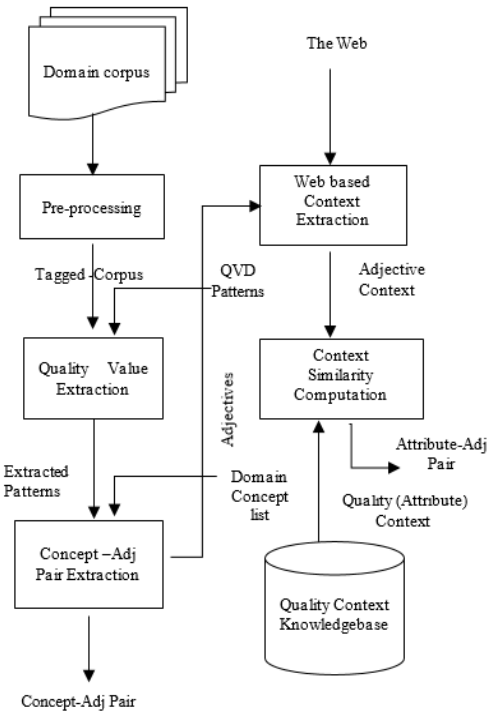

Figure 1. Framework for automatic quality value classification.

3.1. Web Based Quality Context Knowledge Base (QCKB)

The proposed system is based on the assumption that an adjective can be assigned to a quality category if and only if one or more contexts of both are similar. Here in this study, contexts of QA and QV are defined as the various situations in which the quality name can be used. For instance, to assign attribute 'color' to the adjective 'red', it can be directly checked in an enumerated list of colour values. But the classification accuracy will depend on the completeness of the list and also the reliability of sources using which the list is prepared. To avoid this uncertainty, the framework classifies adjectives directly based on their meaning derived from the Web into abstract classes. With strong and impressive support for using the Web as the learning corpus in the literature, we also tend to use the same as reference repository to extract contexts information for various qualities (color, taste, texture etc.,). To mitigate this purpose an online dictionary PyDictionary () has been used. It is a Dictionary Module for Python 2/3 to get meanings, translations, synonyms and Antonyms of words. The dictionary uses Word Net for getting meanings, Google for translations, and thesaurus.com for getting synonyms and antonyms.

For example, contexts of QA 'Colour' as extracted from the Web are:

{'color': {u'Adjective': ['having or capable of producing colors'], u'Verb': ['add color to', 'affect as in thought or feeling', 'modify or bias', 'decorate with colors', 'give a deceptive explanation or excuse for', 'change color, often in an undesired manner'], u'Noun': ['a visual attribute of things that results from the light they emit or transmit or reflect', 'interest and variety and intensity', 'the timbre of a musical sound', 'an outward or token appearance or form that is deliberately misleading', 'any material used for its color', '(physics', 'the appearance of objects (or light sources', 'or brightness']}}

Contexts of quality value 'red' as extracted from the Web are:

{'red': {u'Adjective': ['of a color at the end of the color spectrum (next to orange', 'characterized by violence or bloodshed', '(especially of the face', 'or crimson'], u'Noun': ['red color or pigment; the chromatic color resembling the hue of blood', 'a tributary of the Mississippi River that flows eastward from Texas along the southern boundary of Oklahoma and through Louisiana', 'emotionally charged terms used to refer to extreme radicals or revolutionaries', 'the amount by which the cost of a business exceeds its revenue']}}

Semantic descriptions for each QA are retrieved and stored in a knowledgebase named as Quality Context Knowledge Base (QCKB) as shown in Figure 2. Each tuple in the knowledgebase includes a QA name and its corresponding semantic description. As a part of pre-processing phase context information of QAs is extracted from the web and kept in the knowledgebase.

For illustration: Tuple for QA 'color' in QCKB is as shown below

<'color':['having or capable of producing colors', 'add color to', 'affect as in thought or feeling', 'modify or bias', 'decorate with colors', 'give a deceptive explanation or excuse for', 'change color, often in an undesired manner', 'a visual attribute of things that results from the light they emit or transmit or reflect', 'interest and variety and intensity', 'the timbre of a musical sound', 'an outward or token appearance or form that is deliberately misleading', 'any material used for its color', 'physics', 'the appearance of objects or light sources', 'or brightness']>

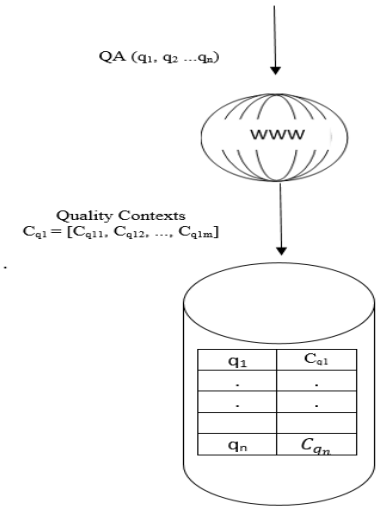

Figure 2. Quality context knowledge base creation.

In the present study the QCKB contains semantic descriptions or contexts for only 5 qualities or QA. The knowledge base can be easily extended to store semantic descriptions for more QAs. In contrast to the existing works in the literature which are mostly dependent on predefined syntactic patterns, the

proposed framework in this paper is first of its kind to deal with the research objective similar to document classification problem. Once descriptions are retrieved and stored in QCKB, a VSM is built which involves creating a vector representation of terms in the semantic description of each QA classes well as the quality value or adjective in the query.

Let QA_class: set of QAs={color, taste, smell, touch/texture, shape}
m: No of distinguished terms in the entire QA_class collection
tf_{ij}: No of occurrences of term m_j in class QA_class_i
df_j: No of classes in which term m_j appeared

$$idf_{j=}\log\frac{N}{df_i} \quad (1)$$

Where *N* is total no of QA_classes

The weighing measure for each feature term m_j in the term vector corresponding to QA_class_i

$$W_{ij}=tf_{ij} * idf_j \quad (2)$$

Let
V: feature weight vector for QA_class_i
$=(W_{i1}, Wi_2\ldots, Wim)$
V_Q: feature weight vector for QV in query Q
$= (wq_1, wq_2\ldots wq_m)$

Then for each QA_class_i, similarity coefficient score between the two vectors V and V_Q is computed using the standard cosine similarity measure and a probabilistic MNB classifier. The QA class associated with pair of vectors which return maximum similarity score is assigned to the corresponding QV.

3.2. Quality Value Extraction from Domain Text

Syntactic patterns have been widely used to extract noun-adjective pair from text as discussed in literature [2, 21]. Similar kinds of patterns have been used in proposed framework as shown in Table 1.

Henceforth throughout the paper these patterns will be referred as QVD patterns. The proposed method of extracting attribute values is similar to those used in the existing literature apart from the difference is that the syntactic patterns used are capable of identifying multiple instances of modifiers present in a phrase for a given concept name.

For instance:

"Fruits are juicy and fleshy"
NN JJ JJ

"The white and pink flowers"
JJ JJ NN

"The big green, white and red house"
JJ JJ JJ JJ NN

The QVD patterns can handle punctuation marks in the phrase. Patterns (P1-P3) are similar to the pattern used in the state of art which takes the general form as shown below

"string1 * string2" (including the double quotes)

Here the wildcard * represents an unspecified single word

for example:

[the red car]
[an expensive gift]

The output of the Quality Value Extraction blocks (i.e., N-Adj pairs) is analysed to identify the domain concepts. Here as framework is dealing with Medicinal Plant Domain (MPD), the domain concepts include plant and plant part name which have been identified through UMLs tagging. They are supplied as input to the subsequent classifier block.

Table 1. Quality Value Detection (QVD) patterns.

Pattern	Example	Pattern	Example
< JJ NN >	[(u'fresh', u'JJ'), (u'leaves', u'NNS')] [(u'yellow-green', u'JJ'), (u'flowers', u'NNS')]	< NN VB JJ |RB JJ >	[(u'plant', u'NN') (u'is ', u'VBZ) (u'living', u'JJ')] [('plant', 'NN') ('is', 'VBZ) ('beautiful', 'JJ')] [(u'rhizomes', u'NNS'), (u'are', u'VBP'), (u'fibrous', u'JJ')]
< JJ NN NN>	[(u'dry', u'JJ'), (u'ginger', u'NN'), (u'root', u'NN')] [(u'pink', u'JJ'), (u'flower', u'NN'), (u'buds', u'NNS')]	< NN VB (JJ |RB JJ) CC (JJ |RB JJ) >	[(u'rhizomes', u'NNS'), (u'are', u'VBP'), (u'juicy',u'JJ'), (u'and', u'CC'), (u'fleshy', u'JJ')]
<JJ CC JJ NN>	[(u'white' ,u 'JJ'), (u'and', u'CC'), (u'pink', u'JJ'), (u'flower', u'NN')]	< NN VB (JJ |RB JJ), (JJ |RB JJ) CC (JJ |RB JJ) >	[('emblic',' NN'), ('is','VBZ'), ('sour','JJ'), (',',','), ('bitter','JJ'), ('and','CC'), ('astringent',' JJ')]
JJ->Adjective NN->Noun (Singular/ Plural) VB->Verb CC->Conjunction RB->Adverb			

3.3. Quality Value Classification

As the key idea behind the research presented in this paper is to overcome the uncertainty in clustering by co-occurrence, the proposed model undertakes the research objective of assigning attribute to adjectives, as classification problem. Experiments have been done with two fundamental classifiers

1. TF-IDF with cosine measure.
2. A Multinomial Naive Bayes classifier.

A detailed description of these algorithms is given in [https://ils.unc.edu/courses/2013_spring/inls509_001/lectures/06- VSM. pdf [18].

The former model computes the similarity coefficient between feature vectors of query adjective and QAs by using cosine similarity measure. Since the length of semantic descriptions of entities of interest i.e., QAs and adjectives are not uniform, we have chosen cosine similarity measure as similarity

measure in this model. The second model is the probabilistic model which computes the similarity coefficient as the probability that the QA will be relevant to the query adjective.

Basically, Naïve Bayes (NB) classifier applies Bayes' hypothesis with strong presumption of independence on selected classifier features. The classifier model can be designed efficiently with relatively little amount of training data. Here, in this experimentation NB classifier has been used as the semantic description of QAs are very short and can be advantageous for NB classifier.

In the following step adjectives associated with those filtered tuples are considered for classification. For adjectives in the pattern, their corresponding meanings are extracted from the web using PyDictionary (), as is done for QAs mentioned earlier. The framework then computes the similarity between the contexts of adjective with the context vectors corresponding to each QA stored in database. Finally, Adjectives are tagged with the QAs with the maximum similarity score. Algorithm 1 depicts the procedural details to classify quality using cosine similarity measure.

Given a set of QAs, Q= {Q_1, Q_2... Q_q} where q is the no. of attributes, a set of feature terms, M= {m_1, m_2, ...,m_n} where n is the length of the feature set.

In a Multinomial classifier model, the class conditional probability of feature term with TF-IDF [18] can be computed as below.

$$P(\frac{m_i}{Q_j})=\frac{w(m_i)|Q_j+\propto}{\sum_{i=1}^{n} w(m_i)|Q_j+\propto v} \quad (3)$$

Where

m_i=a term in the feature set M

$w(m_i)|Q_j$=weight of feature m_i w.r.t quality class Q_j $\sum_{i=1}^{n} W(m_i)|Q_j$=sum of weight of features of feature vector w.r.t. quality class Q_j

α=Laplace smoothing parameter (α: =1)

v: Count of all terms in the training vocabulary set.

Algorithm 1. Quality Value classification using VSM+ Cosine similarity

Input:
Retrieved Adjectives L= [q1, q2, q3… qn]
Weighted features of QAs from vector space model
Output: Classified adjectives
Let Sim_Q = []

1. For each (q) in L
2. Get contexts from the web: C←[c1, c2 ….cm]
3. Generate feature vector F_q for q
4. For each (Q_K) in Q do
5. Compute cosine similarity: Sim_score (F_q,M_{Q_k})
6. Sim_Q←Sim_score
7. End
8. Selected_Quality_class= Class belonging to index of max (Sim_Q)
9. End

In Algorithm 2, methodology to compute the class conditional probability of quality value q using Equation (4) has been elaborated.

$$P(Q_k|q) = \text{argmax} P(Q_k)\prod_{i=1}^{n} p(t_i|Q_k) \quad (4)$$
$$Q_k \in Q$$

Where

t_i: a term from query term set {t_1, t_2, ..., t_n}

$p(Q_k|q)$: probability of quality q w.r.t. quality class Q_k

$p(Q_k)$: prior probability of quality class Q_k

Algorithm 2. Quality Value classification using MNB

Input:
Prior Probability for each Qualitative Attribute (QA)
Weighted features of QAs from vector space model
Query quality value or adjectives L= [q1, q2, q3, …,qn]
Output: Classified Adjectives
1. For each (q) in L do
2. Retrieve context descriptions of input quality value from web
3.Pre-process the description and generate feature terms set T= [t1, t2, ..., tm].
4. Q_prob=[ ]
5. For each (Q_K) in Q do
Total_Feature_Prob=1
For each (ti) in T do
If ti exist in feature set of Q_K do
Total_Feature_Prob= term_weight(ti)*
End
*Class_Prob= Prior Probablity(Q_k) * Total_Feature_Prob*
Q_prob←Class_Prob
End
6. Selected_Quality_class= Class belonging to index of max (Q_prob)
7. End

4. Result Analysis and Evaluation

This section of the paper discusses detail evaluation process of the proposed framework. Performance of the proposed model has been evaluated against structured VSM model by Hartung and Frank [11] as baseline.

4.1. Dataset Description

Apart from comparison to the baseline algorithm, we are mainly more interested to analyse the system performance by using basic classification algorithms in the undertaken research problem. As there is no public text dataset available in Medicinal plant domain, a synthetic dataset has been prepared by collecting text information on general description and medicinal properties of medicinal plants like 'amla', 'turmeric', 'neem', 'curry leaf' , 'aelovera', 'ginger' and 'basil' from different web resources[1,2,3]. There are

[1]http://www.iloveindia.com/indian-herbs/

[2]http://www.indianmedicinalplants.info/Medicinal-Plants

[3]http://www.nhp.gov.in/introduction-and-importance-of-medicinal- plants-and-herbs_hmtl

around 500 sentences in the corpus. UMLS interactive map has been used to identify the various plant parts and qualitative attributes present in the dataset as given in Table 2. These concepts were further filtered to retain only the Concepts of Interest (CI) i.e., quality values belonging to attribute colour, taste, odour, touch/texture and shape. Information content of the test dataset is verified by two domain experts with an Inter Ratter Reliability (IRR) score of 0.78.Various Domain concepts which have been considered in this paper for experimental purpose are 'leaf', 'flower', 'fruit', 'stem', 'rhizome', 'bark' and 'seed'.

Table 2. Description of test dataset.

No of sentences	500
No of qualitative concepts [as per UML tagging]	348
No of Concepts of interest (CI) [as per UML tagging] [taste, colour, texture, odour, shape]	72

4.2. Quality Value Extraction and Classification

Here, in this section experimental evaluation of the proposed framework has been discussed elaborately. Experiments have been conducted to classify the quality values into their abstract quality type with Vector Space Model (VSM) and Multinomial Naive Bayes (MNB) model.

The presented architecture follows a pipelined approach i.e., quality value extraction followed by quality value classification. Hence the overall performance is dependent on the performance of individual stages.

The framework employs a set of predefined patterns i.e., QVD patterns (as mentioned in the earlier part discussion) to extract quality values from domain text. To investigate the retrieval efficiency of QVD patterns, conditional co-occurrence of CI i.e., Quality values belonging to QA types colour, taste, odour, touch, texture and shape and domain concepts (i.e., plant /plant parts) with respect to the NLP patterns given in Table 1 was analysed using Equation (5) proposed in [21]. Table 3 presents a detail analysis of conditional co-occurrence of CI and medicinal plant /plant parts for the test dataset.

$$P(W_i|W_j)= \frac{P(W_j \cap W_i)}{P(W_j)}=\frac{f(W_j \cap W_i)}{f(W_j)} \quad (5)$$

Where W_i: domain concepts (plant/plant parts)
W_j: Concepts of Interest (CI)

The overall <CI, domain concept> pair extraction is found to be 60% by using the QVD patterns. Information retrieval efficiency is attributed to complex and dynamic sentence structure.

For example, sentence regarding medicinal plant 'Neem'[1]:
"*Fruit is one seeded drupe with woody endocarp, greenish yellow when ripe".*

Table 3. Co-occurrence analysis of CI and domain concepts.

Dataset	No of Quality Concepts	Concept of Interest (CI) (%)	Average Conditional Co-occurrence Probability of CI
D1(Amla)	40	25	0.55
D2(Basil)	32	19	0.50
D3(Curry eaves)	30	17	0.86
D4(Ginger)	55	29	0.60
D5(Neem)	45	27	0.70
D6(Turmeric)	70	34	o.27
D7(Aloe Vera)	76	7	0.64

Performance evaluation has been done by computing the classification accuracy of our proposed methodology on106 adjectives and 73 attributes of 'Compositionality Puzzles': Examples from HeiPLAS Development Data set [11] and the synthetic test dataset. Performance is evaluated in terms of precision, recall, fn-score(n=1) and accuracy as shown in Table 4, Figure 3 and Figure 4. 106 adjective of 108 adjectives as listed in this set of HeiPLAS Development Data have been taken for experiment purpose as PyMean() didn't retrieve any meaning for 'dispensible' and 'coars-grained'.

Table 4. Performance analysis of both methodologies.

	HeiPLAS Data set		Test Dataset	
	VSM	MNB+TFIDF	VSM	MNB+TFIDF
Accuracy	0.81	0.83	0.76	0.8
Precision	0.76	0.79	0.73	0.74
Recall	0.79	0.81	0.75	0.79
Fn-Score (n=1)	0.77	0.8	0.74	0.76

VSM with cosine similarity-based classification has attained an accuracy of 81% on HeiPLAS Development Data set and 76% on test data set as shown in Figures 3 and 4. On the contrary, assigning TFIDF based probabilistic weight to feature set in case of MNB classifier has improved the False Positive and False Negative values. Therefore, overall classification accuracy is better in probabilistic approach compared to VSM-based approach. Detailed analysis of obtained result reveals that in certain cases for dataset 1, proposed methodology is capable of capturing the relatedness among adjectives from their semantic definition. For example, adjectives 'brave',' courageous' and 'fearless' are classified into attribute class 'courage' while HeiPLAS dataset labels 'brave', 'courageous' with attribute 'courage' and 'fearless' with attribute 'boldness'.

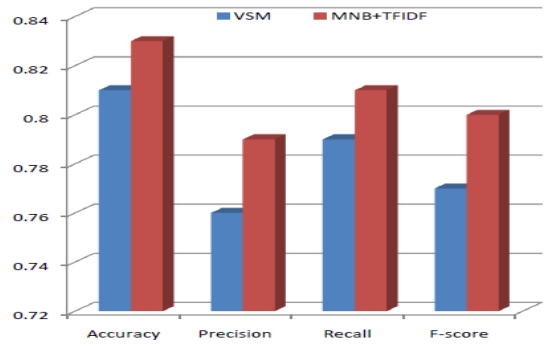

Figure 3. Performance comparison of VSM and MNB with TFIDF on HeiPLAS Development Datase.

Similarly, adjective 'straight' is classified as 'direction' where as in original dataset it is labelled as 'shape'. Unlike the original label 'potency' and 'power' as given in HeiPLAS dataset for adjective 'potent', the proposed model classifies it as 'strength '.To compare the performance of the proposed model with the baseline structured VSM model, experiments were done to retrieve <attribute, adjective> tuples by using patterns A1-A5 and <noun-attribute> tuple using patterns N1-N4 in the test corpus. The A_A (attribute, adjective) patterns and the N_A (noun, attribute) patterns yielded a recall of 0.67and 0.33 respectively. The baseline model reconstructed with this information could retrieve only 0.03% of the concepts of interest that are actually present in the test corpus. The baseline model though doubled the co-occurrence search, still results were not significant as the test dataset had more implicit than explicit attribute-adjective instances. Comparing results of the proposed methodology against the baseline highlights the most important aspects of our model. The decline in performance of the baseline underlines the beauty of our method to infer implicit attribute for noun-adjective phrases.

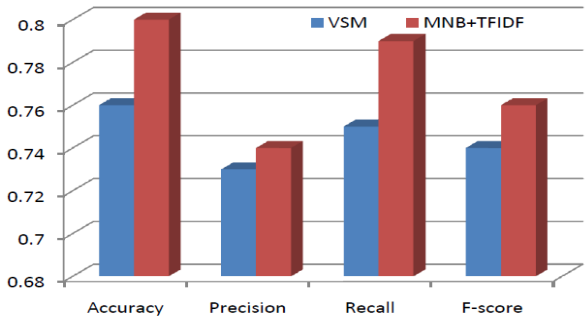

Figure 4. Performance comparison of VSM and MNB with TFIDF on Test dataset on Medicinal plants.

However, though proposed system can perform without human intervention and also involves less computation, the major drawback is that the context descriptions of QAs and attribute values as retrieved from the Web are very short. Hence it is difficult to compute context similarity between the attribute and adjective meaning.
For instance:

evil={u'Adjective': ['morally bad or wrong', 'having the nature of vice', 'having or exerting a malignant influence'], u'Noun': ['morally objectionable behavior', 'that which causes harm or destruction or misfortune', 'the quality of being morally wrong in principle or practice']}

black={u'Adjective': ['being of the achromatic color of maximum darkness; having little or no hue owing to absorption of almost all incident light', 'of or belonging to a racial group especially of sub-Saharan African origin', 'marked by anger or resentment or hostility', 'offering little or no hope', 'stemming from evil characteristics or forces; wicked or dishonorable', '(of events', '(of the face', 'extremely dark', 'harshly ironic or sinister', '(of intelligence operations', 'distributed or sold illicitly', '(used of conduct or character', '(of coffee', 'soiled with dirt or soot'], u'Verb': ['make or become black'], u'Noun': ['the quality or state of the achromatic color of least lightness (bearing the least resemblance to white', 'total absence of light', 'British chemist who identified carbon dioxide and who formulated the concepts of specific heat and latent heat (1728-1799', "popular child actress of the 1930's (born in 1928", '(board games', 'black clothing (worn as a sign of mourning']}}

5. Conclusions

The word 'attribute' has been treated differently like part relations, semantic role labelling etc., by various researchers. In this paper, a framework to extract and classify the qualitative attributes of concepts or entities as an application in medicinal plant domain has been presented. Since the proposed model tries to define real world objects by considering their basic properties or characteristics, this framework can be used to differentiate between entities of two different domains.

Our algorithm can be generalized in order to be applied to various applications which require determining similarity measure between concepts for ontology learning or enrichment which is our ultimate goal. For experimental simplicity, only very few attributes associated with medicinal plant domain concepts have been analysed in this work. In future we are planning to extend our model to include more attributes and investigate the sense ambiguity of adjective-noun meaning. As attribute value acquisition is building block of system, in future we are planning designing more robust patterns to handle complex sentence structure.

References

[1] Acosta O., Aguilar C., and Sierra G., "Using Relational Adjectives for Extracting Hyponyms from Medical Texts," *in Proceedings of the 1st*

International Workshop on Artificial Intelligence and Cognition, Italy, pp. 33-44, 2013.

[2] Almuhareb A. and Poesio M., "Attribute-Based And Value-Based Clustering: An Evaluation," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, pp. 158-165, 2004.

[3] Almuhareb A., "Attributes in Lexical Acquisition," Ph. D, Thesis, University of Essex, 2006.

[4] Bakhshandeh O. and Allen J., "From Adjective Glosses to Attribute Concepts: Learning Different Aspects that an Adjective Can Describe," *in Proceedings of the 11th International Conference on Computational Semantics*, London, pp. 23-33, 2015.

[5] Boleda G., "Automatic Acquisition of Semantic Classes for Adjectives," Ph.D. Dissertation, Pompeu Fabra University, 2006.

[6] Buitelaar P., Cimiano P., and Magnini B., "Ontology Learning from Text. An Overview," *Ontology Learning from Text Methods, Evaluation and Applications*, vol. 123, pp. 3-12, 2005.

[7] Cimiano P., *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications* Springer-Verlag New York, 2006.

[8] Doan P., Arch-int N., and Arch-int S., "A Semantic Framework for Extracting Taxonomic Relations From Text Corpus," *The International Arab Journal of Information Technology*, vol. 17, no. 3, pp. 325-337, 2020.

[9] Gillani S., "From Text Mining to Knowledge Mining: An Integrated Framework of Concept Extraction and Categorization for Domain Ontology," Ph.D. Thesis, Corvinus University of Budapest, 2015.

[10] Guarino N., "Concepts, Attributes and Arbitrary Relations: Some Linguistic and Ontological Criteria for Structuring Knowledge Base," *Data and Knowledge Engineering*, vol. 8, no. 3, pp. 249-261, 1992.

[11] Hartung M. and Frank A., "A Structured Vector Space Model For Hidden Attribute Meaning in Adjective-Noun Phrases," *in Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, pp. 430-438, 2010.

[12] Hartung M. and Frank A., "Semi-Supervised Type-Based Classification of Adjectives: Distinguishing Properties and Relations," *in Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, 2010.

[13] Hartung M., Kaupmann F., Jebbara S., and Cimiano P., "Learning Compositionality Functions on Word Embedding for Modelling Attribute Meaning in Adjective-Noun Phrases," *in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, pp. 54-64, 2017.

[14] Hatzivassiloglou V. and McKeown K., "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning," *in Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, USA, pp. 172-182, 1993.

[15] Kang Y., Haghighi P., and Burstein F., "CFinder: An Intelligent Key Concept Finder from Text for Ontology Development," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4494-4504, 2014.

[16] Lee T., Wang Z., Wang H., and Hwang S., "Attribute Extraction and Scoring: A Probabilistic Approach," *in Proceedings of IEEE 29th International Conference on Data Engineering*, Brisbane, pp. 194-205, 2013.

[17] Liu Z., Chen Y., Dai Y., Guo C., Zhang Z., and Chen X., "Syntactic and Semantic Features Based Relation Extraction in Agriculture Domain," *in Proceedings of International Conference on Web Information Systems and Applications*, Taiyuan, pp. 252-258, 2018.

[18] Mowafy M., Rezk A., and El-bakry H., "An Efficient Classification Model for Unstructured Text Document," *American Journal of Computer Science and Information Technology*, vol. 6, no.1, pp. 1-10, 2018.

[19] Nabila N., Basir N., and Deris M., "Non-Taxonomic Relation Extraction Using Probability Theory," *in Proceedings of World Congress on Engineering and Computer Science*, San Francisco, pp. 287-301, 2017.

[20] Navarro-Almanza R., Juárez-Ramírez R., Licea G., and Castro J., *Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications*, Springer, 2020.

[21] Ong J. and Kliegl R., "Conditional Co-Occurrence Probability Acts like Frequency in Predicting Fixation," *Journal of Eye Movement Research*, vol. 2, no. 1, pp. 1-7, 2008.

[22] Petersen W. and Hellwig O., "Exploring The Value Space of Attributes: Unsupervised Bidirectional Clustering of Adjectives in German," *in Proceedings of COLING the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, pp. 2839-2848, 2016.

[23] Poesio M. and Almuhareb A., "Extracting Concept Descriptions from the Web: the Importance of Attributes and Values," *in Proceedings of the Conference on Ontology Learning and Population*, Amsterdam, pp. 29-44, 2008.

[24] Rios-Alvarado A., Lopez-Arevalo I., and Sosa-Sosa V., "Learning Concept Hierarchies From Textual Resources for Ontologies Construction," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5907-5915, 2013.

[25] Sánchez D., "A Methodology To Learn Ontological Attributes from The Web," *Data and Knowledge Engineering*, vol. 69, no. 6, pp. 573-597, 2010.

[26] Wang C., Fan Y., He X., and Zhou A., "Predicting Hypernym-Hyponym Relations for Chinese Taxonomy Learning," *Knowledge and Information Systems*, vol. 58, no. 3, pp. 585-610, 2019.

[27] Zhao G. and Zhang X., "Domain-Specific Ontology Concept Extraction and Hierarchy Extension," *in Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval*, Bangkok Thailand, pp. 60-64, 2018.

[28] Zhou Y., Zhang L., and Niu S., "The Research of Concept Extraction in Ontology Extension Based on Extended Association Rules," *in Proceedings of IEEE International Conference of Online Analysis and Computing Science*, Chongqing, pp. 111-114, 2016.

Niyati Kumari Behera has completed her M.Tech. in CSE from N.I.T. Rourkela,Odisha.Currently she is persuing her Ph.D in Anna University, Chennai.She has published extensively in peer reviewed journals and international conferences.Her research interest includes Text Mining, NLP and Machine Learning.

Guruvayur Suryanarayanan Mahalakshmi received her Masters in CSE from Anna University, Chennai. She received the Ph.D. during 2009 in the field of Artificial Intelligence. She has authored numerous research articles in Reputed Journals and International Conferences. Presently she is an Associate Professor in Dept. of Computer Science & Engineering, Anna University, Chennai. Her research interests include Machine Learning, Social Networks, Text Mining and Big Data Analytics.