CONTENT BASED AUTOMATIC CLASSIFICATION OF RESEARCH ARTICLES

*Muhammad Faheem Khan, **Aurangzeb Khan, *Shahid Khan , **Aziz Ullah Khan
*EDC, Gandhara University Peshawar, **Institute of Engineering &
Computing Sciences, University of Science & Technology Bannu.
theengineer_soft@yahoo.com, aurangzebb_khan@yahoo.com, Shahid_khan_1@yahoo.com, azizullah50@yahoo.com

ABSTRACT : *Research Article Classification is mainly concern with Document Classification process. Content of the article is used as a "Bag of Word" BOW with term frequency. The vector notation represents the bag-of-words. Various supervised and un-supervised learning techniques are used for classification process like Support Vector Machine (SVM), and KNN, Naïve Bayes etc. The proposed method applied on Research Articles datasets as Training and Testing Datasets which enhance the classification process by using the term frequency - inverse document frequency (TF-IDF) along with features extraction and selection. After finding the concept weight of TF-IDF the Naïve Bayes classifier used to classify the Research articles to the pre-defined category.*

.

KEYWORDS: Document Classification, Feature Selection, TF-IDF, Naïve Bayesian Classifier

I
INTRODUCTION

Automatic classification of Document is the process of allocating documents to a pre-defined category automatically. With the rapid growth and improvement in the Web 2.0 it becomes essential to make the document classification automatically. Proper classification of Research articles, e-documents, online news and reviews, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing techniques to acquire meaningful information by classifying in predefined categories. To minimize the document complexity and handling mechanism the transformation method into document vector from full text version is encouraged, this is one of the pre-processing steps of the document representation. In this step the document is mapped into predefined categories from the contents of the documents. To form a document it is required to represent the term weights vector in order that this term must take place at least one time in a document [2]. Many researchers used the Vector Space Model (VSM) method for categorization of documents. In this method the document is represented as feature term vector where term weights usually contains term-frequencies used in the documents. To easily vectorize the document the most widely used technique i.e. Term Frequency Inverse Document Frequency (TF-IDF) is used for the representation of documents. [3]

Hierarchal categorization of articles and their contents are mostly used for the performance improvement in semantic resources of text classification.[7,8]. Automatic article categorization can be improved with the text classification method. [9,10]. The system that automatically categorizes the research articles will more helpful for researchers and authors. It is very difficult and complex task for inexperienced authors to manually assign category to their articles [11]. Automatic categorization of articles and documents is a supervised learning technique which uses a classifier to categorize the document and articles by related category. Training datasets support the learning process and improves the accuracy in categorization. [12]

For the proper management and handling of research articles it is very important to automatically categorize the research articles according to its contents into pre-defined categories.

In this work a method is proposed To Design, Develop and Evaluate Content Based Automatic Classification of research articles. For the evaluation of proposed method two types of datasets are taken i.e. Training Datasets and Testing Datasets. Using Naïve Bayes Classifier Algorithm along with the Term Frequency Inverse Document Frequency (TF-IDF) techniques for the classification of research articles.

The remaining section of the paper is organized as follows: Section II presents background and Literature Review, in section III proposed method, result and discussion is described in the Section IV, and conclusion and future work is presented in section V.

II BACKGROUND AND RELATED WORK

In literature text classification is broadly studied particularly the Document Classification. Some of the author used supervised learning method for text classification [7]. Many of the other used Unsupervised and Semi Supervised method. [6,13].

Text classification by semi supervised learning is represented in [14] due to the limited categorized text with a maximum level of uncategorized text. They used expectation-maximization (EM) algorithm with a popular model like Naïve Bayes, considering uncategorized data with missing information.

In [3] a new weighting method is introduced where word categorization is performed by a statistical estimation.

The [4] proposed a new method, called class–dependent–feature–weighting by naive Bayes classifier.

Some authors believe that domain ontology is more important in text classification process. According to [15], with the help of hierarchical news ontology, the incoming news can be classified automatically and the system can provide personalized paper to every user related to their profiles. Similarly [16] represents the categorizations of Web documents with the help of weighted terms by ontology.

In [17] documents classification is described by extracting features in a two sets of documents considering only nouns and considering with all Part of speech. After these terms are checked in the WordNet lexical resource along with synonyms as a word features to represent the documents. Similarly [18,19] also used WordNet for the construction of feature vector semantically. Category's contents are extracting by selection of nouns. Sentiment classification is performed in [20] by frequency based feature extraction using SentiWordNet and WordNet.

[21] Use algorithm for extraction of co-occurring terms and sentences and then convert these concept to vector space model (VSM) using domain ontology for the purpose to reduce the feature space. They convert the text firstly to Part of Speech (POS) and then select the frequent terms in a noun form. These terms is then mapped with WordNet Domain knowledge to find the semantic relation.

In this work we use Term Frequency Inverse Document Frequency technique for the classification of research articles. Here we convert text into part of speech (POS) and select only Noun, Verb, Adverb and adjective POS. two types of datasets are considered. One for Training Dataset and second for Testing Dataset. Initially the system is trained with Training dataset and using machine learning technique to extract the contents of research articles and using TF-IDF to categorize and classify the articles to pre-defined categories of training datasets.

III. PROPOSED WORK

Methodology

The process of Electronic Research articles classification requires the supervised learning techniques to convert text into vector model. Various classifiers like Naïve Bayes, Support Vector Machine, and KNN etc. are used for the classification. The Research articles are arranged as Vector model and their contents are represented as words or terms of vector notation. These notations also known as bag-of-words. The bag-of-words are notated with a part of speech tag. The tagged notation contains the both opinion and non-opinion terms. The opinion terms are considered for onward process and non-opinion words are removed which is also called stops-words like "of", "on", "the", "an" etc. the stemming algorithm convert the word of different level to root level. The opinion word contains Noun, Verb, Adverb and Adjective. During this process various features are extracted and then TF-IDF is applied to select the relevant features. Using the Naïve Bayes Classifier to classify the articles to the pre-defined category considering the contents of testing datasets mapping with training datasets.

Following are the steps of content based classification of research articles using TF-IDF and Naïve Bayes Classifier.

1. Select articles and perform processing steps for noise removal.
2. Split the noise free article text into sentences.
3. Remove stop-words from the tokens of the sentences.
4. Apply stemming algorithm to leaves out the word root form.
5. Part of speech tagging of all token using stanford-postagger Part-of-speech tagger.
6. Collect the noun, noun phrases, adjectives, verb and adverb along with their word position in the sentence.
7. extract the features list from key noun phrases
8. Apply TF Algorithm to extract the most frequent terms.
9. Find the TFIDF (Term Frequency and Inverse Document Frequency) to calculate the weight.
10. Calculate the most high relevance sentence using maximum weight or frequency in the document, and give maximum weight.
11. Train the classifier using the new weighted term vector.
12. Calculate similarity between the training and testing documents.
13. Assign the article to the relevant category.

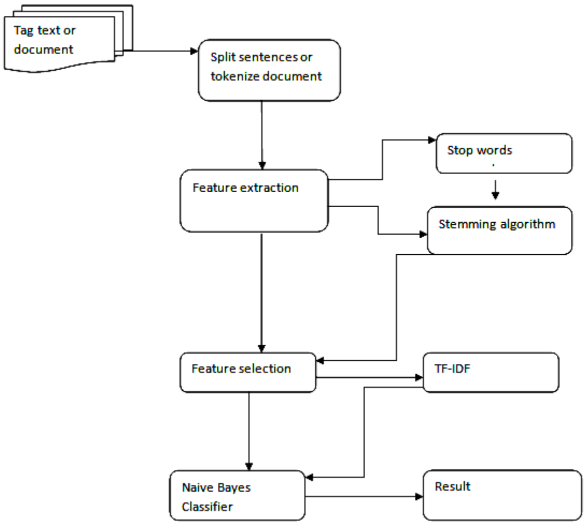


Figure 1: Block diagram of the proposed method

Noise Removal:

As a first step the unwanted text are removed and the content of the articles prepared for onward processing.

Sentence Splitting and Tokenization process:

For classification process the articles is splitted to sentences with unique sentence identification ID using "." As a sentence boundary. After splitting the sentences each sentence again tokenize along with the consideration of the token position.

Remove Stop-Words:

Using stop-words list the non-opinion words /stop-words are removed from the extracted tokens.

Stemming:

Apply the stemming algorithm to find out the stem/root of the tokens.

Part of Speech (POS) tagging:

The standford-postagger Part-of-Speech tagger has been used for tagging each token. In this work we consider only opinion words such as Noun, Adjective, Verb and Adverbs. The standford-postagger assign tag /NN to Noun, /JJ to Adjective, /VB to verb and /RB to Adverb. Table 1 described POS tagging step.

Table 1 Part of Speech Tags

POS-ID	POS-Name	POS-Abbreviation
1	Noun	NN
2	Adjective	JJ
3	Verb	VB
4	Adverb	RB

Feature Extraction:

The opinion words list such as Noun, Verb, Adjective and Adverb are produced along with POS tag. This list is further used for selection of features using Term Frequency (TF).

Frequency base Features Selection Process:

The produced features list consists of many features. For decision making we choose /NN tag and define the threshold frequency. The most frequents words which are greater than threshold value is selected as a frequent term and is considered for further classification steps.

TF-IDF:

TF–IDF, term frequency–inverse document frequency, is a statistical method which calculates the weight of the terms and show that much word is important in the document or corpus.

Naive Bayes classifier:

Naive Bayes classifier is a probabilistic classifier which applies Bayes Theorem and fined the relevance between training and testing datasets with strong independence suppositions.

IV RESULT AND DISCUSSION

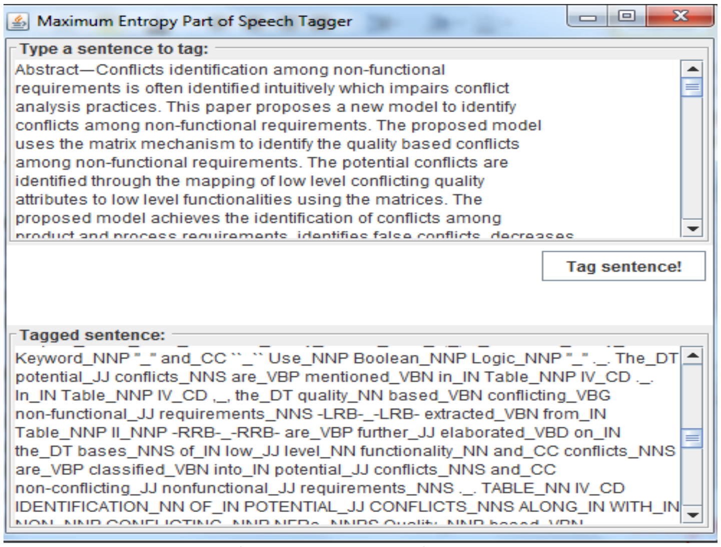
To explain the content based automatic classification of research articles we take an example. For evaluation of our method a dataset of 150 research articles are selected of four different categories in the area of information technology as shown in a below Table II. 75% of these articles are used as a training dataset and 25% articles taken as a testing dataset.

Table II Categories of Research Articles

S.No	Category	No of Articles
1	Software Engineering	40
2	Data Mining	45
3	Computer Graphics	30
4	Network Security	35
	Total	150

The preprocessing steps applied on the training datasets of all four categories, and all the training datasets placed as a knowledge base in part-of-speech tagged form. For evaluation our system the testing datasets processed according to the below steps.

Use standford-postagger Part-of-Speech tagger to tag the testing dataset articles such as shown in below figure 2.

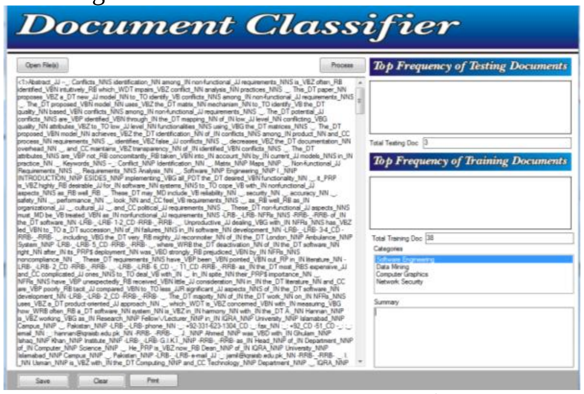


Abstract—Conflicts identification among non-functional requirements is often identified intuitively which impairs conflict analysis practices. This paper proposes a new model to identify conflicts among non-functional requirements. The proposed model uses the matrix mechanism to identify the quality based conflicts among non-functional requirements. The potential conflicts are identified through the mapping of low level conflicting quality attributes to low level functionalities using the matrices. The proposed model achieves the identification of conflicts among product and process requirements, identifies false conflicts, decreases

```
Keyword_NNP "_" and_CC ``_`` Use_NNP Boolean_NNP Logic_NNP "_" ._. The_DT
potential_JJ conflicts_NNS are_VBP mentioned_VBN in_IN Table_NNP IV_CD ._.
In_IN Table_NNP IV_CD ,_, the_DT quality_NN based_VBN conflicting_VBG
non-functional_JJ requirements_NNS -LRB-_-LRB- extracted_VBN from_IN
Table_NNP II_NNP -RRB-_-RRB- are_VBP further_JJ elaborated_VBD on_IN
the_DT bases_NNS of_IN low_JJ level_NN functionality_NN and_CC conflicts_NNS
are_VBP classified_VBN into_IN potential_JJ conflicts_NNS and_CC
non-conflicting_JJ nonfunctional_JJ requirements_NNS ._. TABLE_NN IV_CD
IDENTIFICATION_NN OF_IN POTENTIAL_JJ CONFLICTS_NNS ALONG_IN WITH_IN
NON_NNP CONFLICTING_NNP NFRs_NNPS Quality_NNP based_VBN
```

Figure. 2 standford-postagger Part-of-Speech tagger

The tagged articles are processed by the document classifier for the purpose of features extraction. Testing tagged articles selected in document classifier and it will be classified according to the training datasets. The process is shown in below figure 3.


Document Classifier
Open File(s)
Process
Top Frequency of Testing Documents
Top Frequency of Training Documents
Save
Clear
Print

```
<1>Abstract_JJ --_: Conflicts_NNS identification_NN among_IN non-functional_JJ requirements_NNS is_VBZ often_RB
identified_VBN intuitively_RB which_WDT impairs_VBZ conflict_NN analysis_NN practices_NNS ._. This_DT paper_NN
proposes_VBZ a_DT new_JJ model_NN to_TO identify_VB conflicts_NNS among_IN non-functional_JJ requirements_NNS
._. The_DT proposed_VBN model_NN uses_VBZ the_DT matrix_NN mechanism_NN to_TO identify_VB the_DT
quality_NN based_VBN conflicts_NNS among_IN non-functional_JJ requirements_NNS ._. The_DT potential_JJ
conflicts_NNS are_VBP identified_VBN through_IN the_DT mapping_NN of_IN low_JJ level_NN conflicting_VBG
quality_NN attributes_VBZ to_TO low_JJ level_NN functionalities_NNS using_VBG the_DT matrices_NNS ._. The_DT
proposed_VBN model_NN achieves_VBZ the_DT identification_NN of_IN conflicts_NNS among_IN product_NN and_CC
process_NN requirements_NNS ,_, identifies_VBZ false_JJ conflicts_NNS ,_, decreases_VBZ the_DT documentation_NN
overhead_NN ,_, and_CC maintains_VBZ transparency_NN of_IN identified_VBN conflicts_NNS ._. The_DT
attributes_NNS are_VBP not_RB concomitantly_RB taken_VBN into_IN account_NN by_IN current_JJ models_NNS in_IN
practice_NN ._. Keywords_NNS --_: Conflict_NNP Identification_NN ,_, Matrix_NNP Maps_NNP ,_, Non-functional_JJ
Requirements_NNS ,_, Requirements_NNS Analysis_NN ,_, Software_NNP Engineering_NNP I._NNP
INTRODUCTION_NNP ESIDES_NNP implementing_VBG all_PDT the_DT desired_VBN functionality_NN ,_, it_PRP
is_VBZ highly_RB desirable_JJ for_IN software_NN systems_NNS to_TO cope_VB with_IN nonfunctional_JJ
aspects_NNS as_RB well_RB ._. These_DT may_MD include_VB reliability_NN ,_, security_NN ,_, accuracy_NN ,_,
safety_NN ,_, performance_NN ,_, look_NN and_CC feel_VB requirements_NNS ,_, as_RB well_RB as_IN
organizational_JJ ,_, cultural_JJ ,_, and_CC political_JJ requirements_NNS ._. These_DT non-functional_JJ aspects_NNS
must_MD be_VB treated_VBN as_IN nonfunctional_JJ requirements_NNS -LRB-_-LRB- NFRs_NNS -RRB-_-RRB- of_IN
the_DT software_NN -LRB-_-LRB- 1-2_CD -RRB-_-RRB- ._. Unproductive_JJ dealing_VBG with_IN NFRs_NNS has_VBZ
led_VBN to_TO a_DT succession_NN of_IN failures_NNS in_IN software_NN development_NN -LRB-_-LRB- 3-4_CD -
RRB-_-RRB- ,_, including_VBG the_DT very_RB mighty_JJ reconnoiter_NN of_IN the_DT London_NNP Ambulance_NNP
System_NNP -LRB-_-LRB- 5_CD -RRB-_-RRB- ,_, where_WRB the_DT deactivation_NN of_IN the_DT software_NN
right_NN after_IN its_PRP$ deployment_NN was_VBD strongly_RB prejudiced_VBN by_IN NFRs_NNS
noncompliance_NN ._. These_DT requirements_NNS have_VBP been_VBN pointed_VBN out_RP in_IN literature_NN -
LRB-_-LRB- 2_CD -RRB-_-RRB- ,_, -LRB-_-LRB- 6_CD -_: 11_CD -RRB-_-RRB- as_IN the_DT most_RBS expensive_JJ
and_CC complicated_JJ ones_NNS to_TO deal_VB with_IN ._. In_IN spite_NN their_PRP$ importance_NN ,_,
NFRs_NNS have_VBP unexpectedly_RB received_VBN little_JJ consideration_NN in_IN the_DT literature_NN and_CC
are_VBP poorly_RB tacit_JJ compared_VBN to_TO less_JJR significant_JJ aspects_NNS of_IN the_DT software_NN
development_NN -LRB-_-LRB- 2_CD -RRB-_-RRB- ._. The_DT majority_NN of_IN the_DT work_NN on_IN NFRs_NNS
uses_VBZ a_DT product-oriented_JJ approach_NN ,_, which_WDT is_VBZ concerned_VBN with_IN measuring_VBG
how_WRB often_RB a_DT software_NN system_NN is_VBZ in_IN harmony_NN with_IN the_DT A._NN Hannan_NNP
is_VBZ working_VBG as_IN Research_NNP Fellow\/Lecturer_NNP in_IN IQRA_NNP University_NNP Islamabad_NNP
Campus_NNP ,_, Pakistan_NNP -LRB-_-LRB- phone_NN :_: +92-331-623-1304_CD ;_: fax_NN :_: +92_CD -51_CD -_: ;_:
email_NN :_: hannan@iqraisb.edu.pk_NN -RRB-_-RRB- ._. J._NNP Ahmed_NNP was_VBD with_IN Ghulam_NNP
Ishaq_NNP Khan_NNP Institute_NNP -LRB-_-LRB- G.I.K.I._NNP -RRB-_-RRB- as_IN Head_NNP of_IN Department_NNP
of_IN Computer_NNP Science_NNP ._. He_PRP is_VBZ now_RB Dean_NNP of_IN IQRA_NNP University_NNP
Islamabad_NNP Campus_NNP ,_, Pakistan_NNP -LRB-_-LRB- e-mail_JJ :_: jamil@iqraisb.edu.pk_NN -RRB-_-RRB- ._. I.
_NN Usman_NNP is_VBZ with_IN the_DT Computing_NNP and_CC Technology_NNP Department_NNP ,_, IQRA_NNP
```

Total Testing Doc: 3

Total Training Doc: 38

Categories:
- Software Engineering
- Data Mining
- Computer Graphics
- Network Security

Summary:

Figure. 3 Document Classifier

Here we select three articles from the Software Engineering category as a testing dataset for evaluation of our system. The system extracted the most frequent terms as TF of both training datasets and testing datasets. The most frequent terms which exceed the threshold value are considered for onward classification. In this process method the following features are selected as shown in the below Table III.

Table III Selected Features as a TF of Testing Documents

Word	TF	Category
Requirements	23	1
Non-Functional Requirements	17	1
Software	15	1

Similarly the training documents selected features are extracted as shown in below Table IV.

Table IV Selected Features as a TF of Training Documents

Word	TF	Category
Requirements	29	1
TCP-IP	11	4
2D	11	3

The Term Frequency-Inverse Document Frequency will be calculated for the importance of terms in the articles by using the formula.

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$ ----- {Term Frequency}

In the inverse documents is used to check either the term is common across all document are not.

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$ ---- {Inverse Document Frequency}

To find IDF the total Number of document is divided by the documents that having the TF, then take log of the quotient. Where *D* is the Total Number of Document and *d* is the number of document that contains the *t*. and finally TF-IDF is calculated as

---- {TF-IDF}

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$ After find the TF-IDF the Naïve Bayes classifier has been applied to classify the testing articles to a proper category as shown in the below equation.

$$P(D \mid c_i) = \prod_{j=1}^{n} P(d_j \mid c_i)$$ ---- {Naïve Bayes classifier}

Where D is document and C*i* is defined categories, *i* =1 to *n* and d*j* is the weight of term in document D.

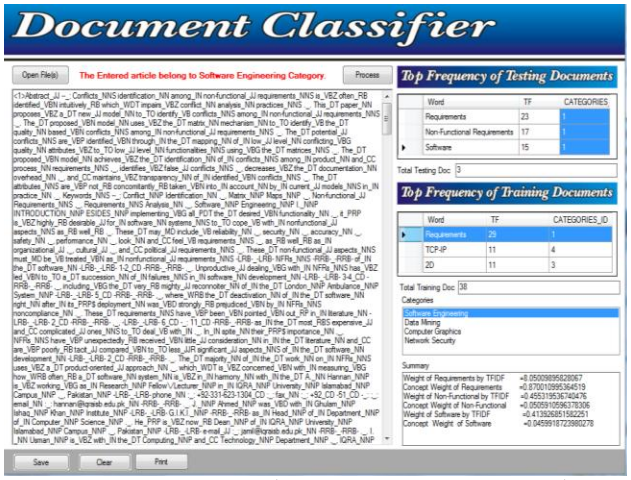
Algorithm: Classification.
Input: Testing documents (TD),
Term set (TS) and Ontology
Output: Categorized training documents

The classification process performed by using Naïve Bayes Classifier and the result shown in below figure 4.

The Term "Requirement" occurs more frequent in both Training and Testing Datasets and the Naïve Bayes Classifier the classified the document to the Software Engineering Category (category Number 1) where the term exists.

V. CONCLUSION AND FUTURE WORK

In this paper Research article Classification is made by using features selection based on the Part-of-speech terms. The TF-IDF method is used for finding Term Frequency Inverse document frequency. Naïve Bayes Classifier is used to the classify the article to pre-define category with the help of training and testing datasets. In future work we will improve the result by using cosine similarity between training documents and testing documents which will enhance the classification result and also the result will compared with other classification methods like MI, SVM and KNN.



The Entered article belong to Software Engineering Category.

```
<1>Abstract_JJ –_: Conflicts_NNS identification_NN among_IN non-functional_JJ requirements_NNS is_VBZ often_RB
identified_VBN intuitively_RB which_WDT impairs_VBZ conflict_NN analysis_NN practices_NNS ._. This_DT paper_NN
proposes_VBZ a_DT new_JJ model_NN to_TO identify_VB conflicts_NNS among_IN non-functional_JJ requirements_NNS
._. The_DT proposed_VBN model_NN uses_VBZ the_DT matrix_NN mechanism_NN to_TO identify_VB the_DT
quality_NN based_VBN conflicts_NNS among_IN non-functional_JJ requirements_NNS ._. The_DT potential_JJ
conflicts_NNS are_VBP identified_VBN through_IN the_DT mapping_NN of_IN low_JJ level_NN conflicting_VBG
quality_NN attributes_VBZ to_TO low_JJ level_NN functionalities_NNS using_VBG the_DT matrices_NNS ._. The_DT
proposed_VBN model_NN achieves_VBZ the_DT identification_NN of_IN conflicts_NNS among_IN product_NN and_CC
process_NN requirements_NNS ,_, identifies_VBZ false_JJ conflicts_NNS ,_, decreases_VBZ the_DT documentation_NN
overhead_NN ,_, and_CC maintains_VBZ transparency_NN of_IN identified_VBN conflicts_NNS ._. The_DT
attributes_NNS are_VBP not_RB concomitantly_RB taken_VBN into_IN account_NN by_IN current_JJ models_NNS in_IN
practice_NN ._. Keywords_NNS –_: Conflict_NNP Identification_NN ,_, Matrix_NNP Maps_NNP ,_, Non-functional_JJ
Requirements_NNS ,_, Requirements_NNS Analysis_NN ,_, Software_NNP Engineering_NNP I._NNP
INTRODUCTION_NNP ESIDES_NNP implementing_VBG all_PDT the_DT desired_VBN functionality_NN ,_, it_PRP
is_VBZ highly_RB desirable_JJ for_IN software_NN systems_NNS to_TO cope_VB with_IN nonfunctional_JJ
aspects_NNS as_RB well_RB ,_, These_DT may_MD include_VB reliability_NN ,_, security_NN ,_, accuracy_NN ,_,
safety_NN ,_, perfomance_NN ,_, look_NN and_CC feel_VB requirements_NNS ,_, as_RB well_RB as_IN
organizational_JJ ,_, cultural_JJ ,_, and_CC political_JJ requirements_NNS ._. These_DT non-functional_JJ aspects_NNS
must_MD be_VB treated_VBN as_IN nonfunctional_JJ requirements_NNS -LRB-_-LRB- NFRs_NNS -RRB-_-RRB- of_IN
the_DT software_NN -LRB-_-LRB- 1-2_CD -RRB-_-RRB- ._. Unproductive_JJ dealing_VBG with_IN NFRs_NNS has_VBZ
led_VBN to_TO a_DT succession_NN of_IN failures_NNS in_IN software_NN development_NN -LRB-_-LRB- 3-4_CD -
RRB-_-RRB- ,_, including_VBG the_DT very_RB mighty_JJ reconnoiter_NN of_IN the_DT London_NNP Ambulance_NNP
System_NNP -LRB-_-LRB- 5_CD -RRB-_-RRB- ,_, where_WRB the_DT deactivation_NN of_IN the_DT software_NN
right_NN after_IN its_PRP$ deployment_NN was_VBD strongly_RB prejudiced_VBN by_IN NFRs_NNS
noncompliance_NN ._. These_DT requirements_NNS have_VBP been_VBN pointed_VBN out_RP in_IN literature_NN -
LRB-_-LRB- 2_CD -RRB-_-RRB- ,_, -LRB-_-LRB- 6_CD -_: 11_CD -RRB-_-RRB- as_IN the_DT most_RBS expensive_JJ
and_CC complicated_JJ ones_NNS to_TO deal_VB with_IN ._. In_IN spite_NN their_PRP$ importance_NN ,_,
NFRs_NNS have_VBP unexpectedly_RB received_VBN little_JJ consideration_NN in_IN the_DT literature_NN and_CC
are_VBP poorly_RB tacit_JJ compared_VBN to_TO less_JJR significant_JJ aspects_NNS of_IN the_DT software_NN
development_NN -LRB-_-LRB- 2_CD -RRB-_-RRB- ._. The_DT majority_NN of_IN the_DT work_NN on_IN NFRs_NNS
uses_VBZ a_DT product-oriented_JJ approach_NN ,_, which_WDT is_VBZ concerned_VBN with_IN measuring_VBG
how_WRB often_RB a_DT software_NN system_NN is_VBZ in_IN harmony_NN with_IN the_DT A._NN Hannan_NNP
is_VBZ working_VBG as_IN Research_NNP Fellow\/Lecturer_NNP in_IN IQRA_NNP University_NNP Islamabad_NNP
Campus_NNP ,_, Pakistan_NNP -LRB-_-LRB- phone_NN :_: +92-331-623-1304_CD ;_: fax_NN :_: +92_CD -51_CD -_: ;_:
email_NN :_: hannan@iqraisb.edu.pk_NN -RRB-_-RRB- ._. J._NNP Ahmed_NNP was_VBD with_IN Ghulam_NNP
Ishaq_NNP Khan_NNP Institute_NNP -LRB-_-LRB- G.I.K.I._NNP -RRB-_-RRB- as_IN Head_NNP of_IN Department_NNP
of_IN Computer_NNP Science_NNP ._. He_PRP is_VBZ now_RB Dean_NNP of_IN IQRA_NNP University_NNP
Islamabad_NNP Campus_NNP ,_, Pakistan_NNP -LRB-_-LRB- e-mail_JJ :_: jamil@iqraisb.edu.pk_NN -RRB-_-RRB- ._. I.
_NN Usman_NNP is_VBZ with_IN the_DT Computing_NNP and_CC Technology_NNP Department_NNP ._. IQRA_NNP
```

Top Frequency of Testing Documents

Word	TF	CATEGORIES
Requirements	23	1
Non-Functional Requirements	17	1
Software	15	1

Total Testing Doc: 3

Top Frequency of Training Documents

Word	TF	CATEGORIES_ID
Requirements	29	1
TCP-IP	11	4
2D	11	3

Total Training Doc: 38

Categories
- Software Engineering
- Data Mining
- Computer Graphics
- Network Security

Summary

```
Weight of Requirements by TFIDF      =8.05009895828067
Concept Weight of Requirements       =0.870010995364519
Weight of Non-Functional by TFIDF    =0.455319536740476
Concept Weight of Non-Functional     =0.0505910596378306
Weight of Software by TFIDF          =0.413926851582251
Concept  Weight  of Software         =0.0459918723980278
```

Figure. 3 Article Classification using Naïve Bayes Classifier

Description	Values
Weight of Requirements by TFIDF	8.05009895828067
Concept Weight of Requirements	0.870010995364519
Weight of Non-Functional Requirements by TFIDF	0.455319536740476
Concept Weight of Non-Functional Requirements	0.0505910596378306
Weight of Software by TFIDF	0.413926851582251
Concept Weight of Software	0.0459918723980278

VI. REFERENCES

[1] Aurangzeb khan, Baharum Baharudin, Khairullah khan, "*Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification*" Second International Conference on Computer Engineering and Applications, 2010 Crown Copyright DOI 10.1109/ICCEA.2010.228

[2] Jingnian Chen, Houkuan, Shengfeng Tian, Youli Qu, *"Feature Selection for Text Classification with Naïve Bayes"*, Expert system with applications, 2009, pp 5432-5435.

[3] P. Sccuy, G.W.Mineanu *"Beyoned TFIDF weighting for text Categorization in the Vector Space Model"*, 2003.

[4] Hiroshi Ogura, Hiromi Amano, Masato Kondo "Feature selection with a measure of deviations from Poisson in text categorization" Expert Systems with Applications 36, - pp 6826–6832, 2009.

[5] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem- Aghaee, Mohammad Ehsan Basiri *"Text feature selection using ant colony optimization"*, Expert Systems with Applications 36 pp.6843–6853, 2009.

[6] E. Youn, M. K. Jeong , *"Class dependent feature scaling method using naive Bayes classifier for text datamining "* Pattern recognition Letters , 30 (5), 477-485,2009.

[7] Zeno Gantner,Lars Schmidt-Thieme,*"Automatic Content-based Categorization of Wikipedia Articles"* Proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, pages 32–37, Suntec, Singapore, 7 August 2009. ACL and AFNLP

[8] Somnath Banerjee. *"Boosting inductive transfer for text classification using Wikipedia".* In ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications, Washington, DC USA., page 148-153. IEEE Computer Society 2007.

[9] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. *"Liblinear: A library for large linear classification".* Journal of Machine Learning Research, Volume 9, 6/1/2008, Pages 1871-1874

[10] Linyun Fu, Haofen Wang, Haiping Zhu, Huajie Zhang, Yang Wang, and Yong Yu. 2007. *"Making more Wikipedians: Facilitating semantics reuse for wikipedia authoring".* In ISWC/ASWC 2007, pp.128-141.

[11] Evgeniy Gabrilovich and Shaul Markovitch. 2007. *"Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization".* "Journal of Machine Learning Research, 8(Oct):2297-2345, 2007.

[12] D. Merkl, "*Text data mining*". In: A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text. Marcel Dekker, New York, 1998, pp.396-207.

[13] Bin Zhang, Mari Ostendorf,"*Semi-Supervised Learning For Text Classification Using Feature affinity Regularization*" IEEE 2012, ICASSP 2012: 5129-5132.

[14] Kamal Nigam, Andrew Mccallum, and Tom Mitchell, "*Semisupervised text classification using EM,*" in Semi Supervised Learning, pp. 33–56. MIT Press, 2006.

[15] Lena Tenenboim, Bracha Shapira, Peretz Shoval "*Ontology-Based Classification Of News In An Electronic Newspaper*" International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008.

[16] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang *"Ontology- Based Automatic Classification and Ranking for Web Documents"* Fourth International Conference on Fuzzy Systems and Knowledge Discovery -FSKD -2007.

[17] T. Masuyama and H. Nakagawa ,"*Cascaded Feature Selection in SVMs Text Categorization*" , LNCS 2588, pp. 588-591, 2003.

[18] D. Lewis, "*Feature selection and feature extraction for text categorization*" . In Proceedings of a Workshop on Speech and Natural Language, 1992, pp. 212-217, San Mateo, CA: Morgan Kaufmann.

[19] D. Lewis, "*An evaluation of phrasal and clustered representations on a text categorization task*" . In Croft et. al. (Ed.), Proceedings of SIGIR- 95, 15th CM International, 1992

[20] Abdul Wahab, Aurangzeb Khan, Muhammad Faheem Khan And Shahid Khan, "*Important Feature Extraction During Sentiment Analysis*" Sci.Int(Lahore),26(2),959-964,2014

[21]. Aurangzeb khan, Baharum Baharudin, Khairullah khan "*Semantic Based Features Selection and WeightingMethod for Text Classification*" Second International Conference on Computer Engineering and Applications, pp.850-855, IEEE 2010.