# What works Better? A study of classifying the requiremnets:

In this Paper two mainly goals were target first is classifying the FR and NFR and second classification of NFR into categories. For the first problem, preprocessing before applying the ML is proposed. As a first step POS tagging, Entity tagging and temporal tagging used. POS tagging assign the part of speech to each word and each requirement. Entity tagging replaced all user and product by name as PRODUCT and USER. Temporal logic used for tag a sample set of NFR and normalized the data set with temporal expression. SUTIME a rule based temporal tagger used with TIMEX3 setting. After that feature co-occurrence and regular expression used increase the weight of influential words used in NFR. J.48 DT used for the classifying the FR and NFR. This preprocessing improved the accuracy from 89.92% to 95.04%.

For the second goal i.e. categorization or classification of NFR Topic Modeling unsupervised algorithm LDA and BTM applied. For generation of the topic Clustering with hierarchical Agglometric Algorithm, k-means, Hybrid clustering and k-means also applied. In the last author implemented BNB for the same purpose. The result showed BNB worked better out of clustering, k-means, LDA, BTM. Preprocessing steps improved precision and recall doubled for LDA and BNB. Also BTM works better than LDA for the short text.

# Mining Association between Quality Concern and Functional Requirements:

The goal of the paper is to identify quality concern related to the specific functional components.

For this study Positive Train Control, Electronic Health Records, and Medical Infusion Pumps domain documents dataset used. As a solution, Holistic Requirement Mining approach proposed to provide the analysis support to Engineer for deeper understanding of existing system requirements and improve an existing set of system requirements, or simultaneously mine functional requirements and quality concerns from a domain repository. HRM process has eight steps:

1. Collect large set of domain documents
2. Build domain model of these documents. Automated approach did not satisfy the result in previous research. That is why now manually constructed domain models used as input to this step.
3. Query Expansion by expending the term
4. Search the domain document for the information retrieval
5. Initial search seeds fir quality concerns
6. Search retrieved information or components for quality concerns
7. Categorize the association knowledge
8. Extract requirements from the collected info.

For the query expansion search existing descriptor used to formulate the query. Further expansion is done by applying the section extraction and removal of stopper, POS analysis on relevant text and frequency of noun and pronoun. The quality concern seeds terms generated by the google simple search e.g. security term. After processing searched terms were passed to search engine and relevant doc were retrieved. VSM used to identify the most relevant chunk in each document. Domain Term Frequency (DTF) and Domain Specificity (DS) metrics used to identify and rank candidate term and phrases for inclusion in search term. Total 10 types/ categories of knowledge can be obtained from the interdependency descriptions extracted from domain documents. All of these 10 categories can potentially provide assistance to the engineers in understanding the domain and the existing system requirement.

++++Domain document: project proposals, requirement specifications,architectural designs, implementation plans, product brochures, domain surveys, presentation material, regulatory

codes, and other associated literature.

# Requirement identification with Word2vec and Tensor flow

This paper is from the RE data challenge track, with the focus of the identification of the requirements mainly focusing on the security related requirements. Two datasets with already classification of NFR and FR and Security related and non-security related are used. This classification done by using Multinomial Naïve Bayes. In NFR dataset quality related information was added too. Tensorflow is used to implement the CNN. First data filtration is done by removing the stop words and adding space to the punctuation. As a next step whole documents is converted into the vocabulary index matrix and used as an input to embedding layer. The CNN classifier equipped with Word2Vec embeddings provided an F1-score lift of over 7% over baseline methods for the SecReq dataset and around for the NFR dataset 5.5%. As a future direction quality attribute will consider for the classification of NFR. Word2Vec embedding for requirements representations in automated trace recovery problems.

# Automated Extraction of Conceptual Models from user stories via NLP.

In this paper, a tool named as Visaul Naraator is proposed. The tool takes user stories as an input and as an output generated the conceptual model. For generating the conceptual model, they choose total 11 heuristic derived from NLP. For the future work author mentioned to resolve inconsistencies, redundancies between requirements.

1. Mining Twitter Feeds for Software User

Requirements

In this paper, total 4000 randomly selected tweets were gathered from the ten different software. These tweets were classified in req, bug, and other categories. The search of related tweets was performed differently e.g. used @ windown10 rather than #window10. It helped to overcome the noise problem. First manually tweets were labelled with the re, bug, and spam. This manual analysis showed that 51 % of these tweets were technically informative (27% bug reports and 24% user requirements), while the other 49% were spam and miscellaneous. For the automatic classification SVM and Naïve Bayes algorithm with sentiments, sentence stemming, stop word removal implemented. The results showed that other features used as supplement attributes were irrelevant and the main contribution was only Tweet textual content. The next phase explained in the paper was tweets extractive summary. To enhance the quality of the generated summaries, English stop-words were excluded from our frequency analysis. Stemming was also applied to minimize the redundancy imposed by the usage of different variations of words (e.g., show, showing, shown, and shows). SumBasic produced the best results in comparison to than hybrid TF.IDF and TF in summarizing

the common concerns found in software-relevant tweets. In the future direction, more data collection and analysis with more advanced software with the tool implementation were mentioned.

In this paper oversampling/ under sampling has not discussed. It is obvious from the results that informative were 27 and 24 percent while the spam were classified as 49 percent.

1. Automatic Classification of Non-Functional Requirements from Augmented App User Reviews