Journal of Engineering and Applied Sciences 13 (Special Issue 1): 2407-2412, 2018
ISSN: 1816-949X

Semantic Based Short Messages Classification with Topic Modeling Support

[1]Ghaidaa A. Al-Sultany and [2]Raghad M. Hatim
[1]Department of Information Network,[2]Department of Software,
Babylon University, Babel, Iraq

Abstract: Typically, lexical semantics have focused on how to correlate the meaning of the lexical item with its syntax with respect to the language construction. Due to the hidden semantic structures in text content, Latent Semantic Analysis (LSA) as one of the topic modeling algorithms has shown an effectiveness to treat with the text noisiness, high dimensional issues semantically. It can distinguish the most informative and discriminative features from a collection of text. In this study, the issue of dimensional sparse of the short messages features was avoided through enriching the message's text with linguistic semantics. The enriched features have been fetched to the LSA algorithm to produce the features vector transformation and enhancing the classification process. The experiments of the research have shown very promising results in comparison to the most popular machine learning methods. The classification performance in terms of the evaluation metrics has been discussed and compared against the results without the proposed enriching.

Key words: Short messages, fuzzy rough sets, topic modeling, semantic lexicon, enriching, discussed

INTRODUCTION

Text messages are one of the most popular tools that has employed in communication. These messages are often misapplied by reckless people to perpetrate offenses such as the fraud. This resulted in insecurity by appearing so-named as junk short messages (Holzinger *et al*., 2014; Sheu *et al*., 2009). There are a lot of obstacles for generating efficient short messages classification regulations such as the robustness, the effectiveness, the reliability, etc. However, the major challenges faced in the domain are those posed by the natural short messages themselves such as abbreviations, idioms and small size and nosiness in short messages.

Topic modeling are statistical methods which have been proven effectiveness to deal with noisiness and high dimensional that found in text data through hidden semantic structures in text content. There are many topics modeling algorithms can be used to extract discriminative features without affected on useful data through by focusing on semantic context of text (Blei, 2012; Yin *et al*., 2011), for example, of significant topics modeling methods is Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), hierarchical dirichlet process, explicit semantic analysis, factorization. Latent Semantic Analysis (LSA) is an novel and common method that has been used in text mining (Homayouni *et al*., 2004; Price and Zukas, 2005). It is factorized relations among features and topics included in unstructured set of texts. It has the ability to connect semantically relevant features that are underlying in a text. LSA depends on its work on Singular Value Decomposition (SVD) (Ding *et al*., 2011) to detect patterns among features and topics and determine the relationships among texts. Furthermore, linguistic semantic used to improve the available information about words in texts (Agrawal and Agrawal, 2014). Lexical semantic is a subfield of linguistic semantics which have been proven importance and effectiveness in text processing. The items of analysis in lexical semantics are lexical items that contain not only words but also sub-words or sub-parts such as affixes and even phrases and compound words. Lexical item forms the index of words in a language the lexicon. Lexical semantics is searching at how to link the meaning of the lexical items with the construction of the language or syntax. This is indicated to as syntax-semantic interface (Holzinger *et al*., 2014).

Lexical items also indicated as syntactic atoms, can use alone such as portions of complex words and root words or attach to other parts such as suffixes and prefixes. The first are named free morphemes and the second is bound morphemes (Sheu *et ai*., 2009). It falls into a small domain of meanings (semantic fields) and can merge with each other to create new concepts. Lexical semantics detects whether the concept of a lexical item is proved through searching at its neighbors in the semantic

Corresponding Author: Ghaidaa A. Al-Sultany, Department of Information Network, Babylon University, Babel, Iraq

net (words it exists in natural sentences) or whether the concept is locally found in the lexical item. There are many sources for semantic analysis such as latent semantic analysis and WordNet, in English; WordNet is a dictionary from stanford for the English language depends on psycholinguistics researches and sophisticated at the University of Princeton (Holzinger *et al*., 2014; Sheu *et al*., 2009). It was assumed as a data-processing source that covers lexico-semantic labels named synsets. So, from the information that provide lexicon semantic about the words may be enhance performance the LSA through focused on semantic relation among features and this information.

In this study enriching the message's text with linguistic semantics has been proposed to overcome the difficulty of dimensional sparse in short texts messages. In addition, topic based LSA algorithm has been implemented to provide the features vector transformation and optimizing the classification process based on Quick rules based fuzzy rough set classifier.

Literature review

Semantic based text analysis: Semantic similarity does a significant function in information retrieval, natural language processing, texts summarization, texts clustering and texts categorization. WordNet is the most important lexical English database that commonly used to cover several special words from each subject concern to their stems. WordNet 2.0 includes 155327 words, 207016 of word-sense and 117597 senses. It set of nouns, adjectives, adverbs and verbs to group of synonyms named synsets. Each synset is orderly to senses, granting therefore synonyms for every word. The enforcement of integrating th semantic features based on WordNet lexical semantic database (Miller and Charles, 1991) to enhance the precision of texts clustering methods such as Dave *et al.* (2003) utilized synonyms as features to documents representations and thus clustering. WordNet senses disambiguation was not executed and WordNet synonyms really reduced clustering execution. As result as Hotho *et al.* employed WordNet for documents clustering to words senses disambiguation to enhance the clustering implementation.

Several semantic similarity measurements have been suggested. In generally, these measurements can be classified to four types: measurement depends on path length, measurement depends on information content, measurement depends on feature and hybrid measurement. An overview of these measurements in Meng *et al.* (2013). Rada *et al.* (1989) suggested a system depended on the MeSH ontology to enhanced texts retrieval. The system will be calculated semantic resemblance directly from where the count of ontology is a measurement of the conceptual space among words. Edges among words in hierarchy. Their hypothesis of this system is that the number of edges between terms in on.

Topic modeling based text classification: The prosperity of machine learning methods relies on chosen appropriate features collection for the given problem. A known trouble in text classification is the large dimensional of features space. Topic modeling techniques have been proven effectiveness in texts classification problems because it capability to deal with the nature of the texts that contain noisiness and high dimensional problems. suggested applying Latent Dirichlet Allocation (LDA) as dimensional reduction technique to enhanced clustering documents by K_means, so it's have been detected the elementary clustering centers by determining the typical latent topics elicited by LDA. The performance of LDA_K-means is applied on the 20 Newsgroups data collections. It has been show that LDA_K-means can greatly improve the clustering impact in contrast to clustering based on random initialization of K-means. A short messages filtering uses information gain (Gomez *et al*., 2006; Sohn *et al*., 2009) and mutual information (Deng and Peng, 2006) that are commonly reasonable techniques in texts classification. The researchers by Guan *et al.* (2012) utilized Support Vector Machines (SVM) with information gain as features selection that depends on the messages content. This framework has suffered from biased towards features that have high values.

Data representation and preparing

Dataset: The English junk short messages corpus V.1 2012 was used in this research. A collection of 421 random short messages that labeled as clean or junk message was implementing for system testing and evaluating the system. The selected dataset has subgroup of 372 messages with 'ham-cleans message's type and 49 messages with 'spam message's type. The average of the characters count per message is 4.44 characters and the rate count of words for every text message is 15.725. The data collection includes single text message for each line. Every line contains two columns: class label (junk or not junk) and raw text message column.

Tokenization process: The messages text collection is passed to the tokenization process after removing all the delimiters, stop words, abbreviations, prepossessions and informal expressions. The tokenization process splits the collection text into individual terms (unigrams), since, the

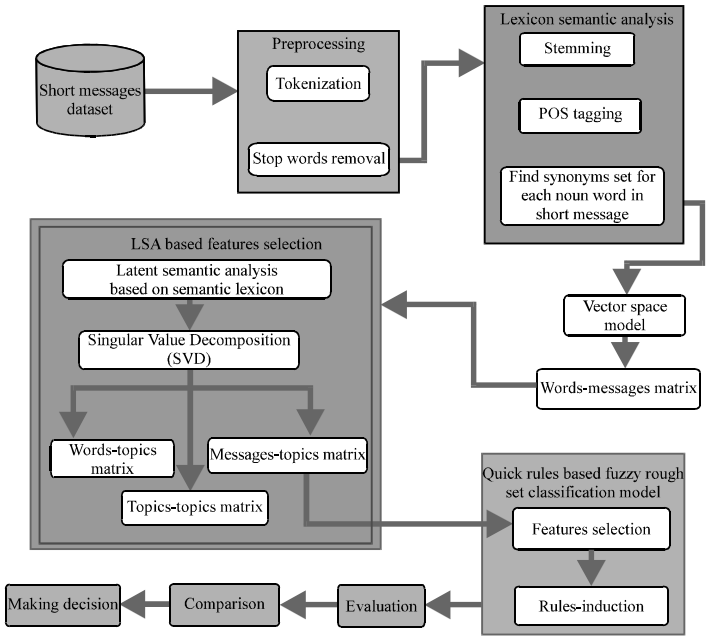

Fig. 1: Features selection using LSA

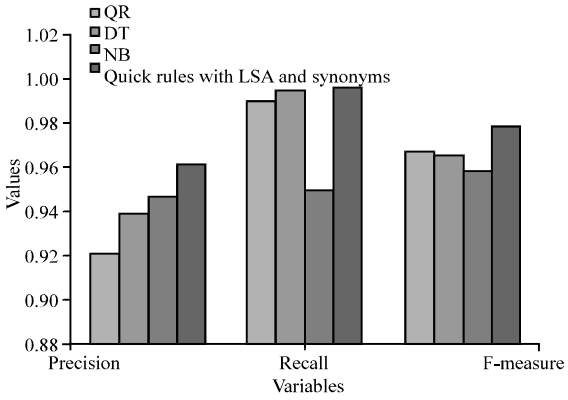

Fig. 2: The chart of the system performance

individual terms may have has the capability to determine the label of the entire text message as shown in Fig. 1 and 2 the system structure.

Stemming process: The purpose of a stemming algorithm is to extract the stem or radix of a word which it is morphological root through deleting the affixes and loading the lexical or grammatical information around a given word. In any case, these affixes do not change the meaning of the word where is connected to as the semantic informality has been established specially in languages construction. In the research Lovins Stemmer have been employed which based on truncating approach to extract the word's stems.

Part of speech tagging: Part of speech tagging was required in our research as the NOUN part of speech was extracted from the text messages as the main terms features. The output noun stems are extracted into a menu of POS tags. This part of a speech indicates to terms which are employed to name things, persons, animals, ideas, places or events. The stanford tagger is employed in our system which it represents a high precision POS tagging technique.

Semantic based features selection

Semantic lexicon: The sparsity and the short length of tweets was address by augmenting their polysemous term with multiple word senses from the semantic lexicon 'WordNet' for increasing the dimensionality of the feature space. Specifically, small subsets of semantic features (core semantics) with the help of information from WordNet have been retrieved. The core semantics can affect positively on representing the main theme of the extracted topics in the text (Fodeh *et al*., 2011). In this research, the synonymy part of the core semantic for nouns is considered and aggregated with the origin text to enlarge the messages volumes increase the accuracy of the classifier of the extracted tokens with their semantic synonyms and stems' root are shown in Table 1.

Vector space model: Transformation every short message to weighted features vector where the features refer to the given terms (message's tokens) and their significance

Table 1: Message's tokens

Token	Stemming	Synsets. Noun	Token	Stemming	Synsets. Noun
Parking	Park	Parkland, commons	Dinners	Dinner	Dinner party
Failing	Fail	-	Stop	Stop	Halt,stoppage
Wants	Want	Neediness, lack	Texts	Text	Textbook, text edition
Plays	Play	-	Renewal	Renew	-
Details	Detail	Item, point	Right	Right	Right hand, rightfulness
Sitting	Sit	-	Bawling	Bawl	-

Table 2: The terms frequencies

Token	M1	M2	M3	M4	M7
Letter	0.693147	0	0	0	0.693147
Particular	0.693147	0	0	0	0
Point	0.693147	0.693147	0	0	0
Problem	1.098612	0	0	0	0
Renewal	0.693147	0	0	0	0
Feeling	0	0.693147	0	0	0
Field	0	0.693147	0.693147	0	0
Hand	0	0.693147	0.693147	0	0
Need	0	0.693147	0	0	0
Neediness	0	0.693147	0	0	0
Rightfulness	0	0.693147	0	0	0

Table 3: Messages-topics matrix

No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
M1	0.141662	-0.028990	0.978305	-0.135840	-0.001010
M2	0.965963	0.212381	-0.139960	-0.036520	0.018908
M3	0.988733	-0.011340	0.129440	0.059897	-0.036900
M4	0.147806	0.005269	0.021007	-0.018030	0.002370
M5	0.023722	-0.006010	0.029926	-0.031830	-0.163850

(weights). The weights were represented using term's frequencies TF. In the study vector space model using frequency each word in all messages in the shape <Word, frequency>. Each value in Table 2 denote the weight for each word in each message according to the TF measure which compute the log of frequency times for each word appear in a given message as log (1+ frequency W_iM_j). Adding log is to dampen the importance of terms which have a high frequency. Each row in words-messages matrix is a vector corresponding to a word, giving its relationship to each message.

Latent semantic analysis for features selection: LSA has adopted for the text mining in particular for features selection as it successfully dealing with text and mining the significant features of. It depends on Singular Value Decomposition (SVD) through focusing on the context of text and semantic relation among the terms. In this research LSA was implemented on the text messages for selecting the textual features as the form of three matrix, features-topics, topics-topics and messages-topics. The messages-topics was selected to be the inputs for classification process to distinguish or group similar messages which can be used in features selection step in quick rule classifier as shown in Table 3 that contain sample of messages-topics. For example, M1, M5 are assigned to topic 3 and M 2-4 are assigned to topic 1. Thus, they will be depended on the Topic 1, Topic 3 for classifying the messages. Dimensionality reduction exemplary step in the text mining that convert the data representation to a smaller, compact and more predictive. The novel space is simpler to process because of its size and to load the more significant parts of information necessary to differentiate among short messages. In features selection methods, a subset of main features is chosen and just the chosen features are utilized for training and testing for the classification algorithms. The deleted features are not employed in the computations anymore. Adding semantic features to features space may be used to enhanced dimensional reduction method through improving the quality of the topics where they will be included more similar features by using the synonyms relationship.

Classification mode-quick rules: In the system have been suggested quick rules based fuzzy rough as classification technique. It relies on the principle of rules induction for features selection. Quick reduct technique is an efficient method for computing Reduct. Quick reduct algorithm tries to compute a reduct without exhaustively creating all subsets it will take messages-topics matrix as data to compute reduct set according to the following Calculating the similarity matrix for each topic based on Eq. 1:

$$\text{Similarity measure} = 1-\frac{(\text{abs}(a(x)-a(y))}{\text{abs}(a\_{max}-a\_{min}))} \quad (1)$$

Table 4: The system evaluation against the known techniques

Algorithms	Precision	Recall	F-measure	Class
Quick rules	0.921	0.990	0.967	Not junk
	0.857	0.500	0.632	Junk
Decision tree	0.939	0.995	0.966	Not junk
	0.926	0.510	0.658	Junk
Naive bayes	0.946	0.950	0.958	Not junk
	0.800	0.898	0.846	Junk
Quick rules with	0.961	0.997	0.979	Not junk
LSA and S lexicon	0.971	0.694	0.810	junk

where, a(x), a(y) are attribute values in each topic, $a_{\_max}$ $a_{\_min}$ are the maximum and Minimum values, respectively. Computing the lower approximation based on the above similarity matrix for each topic based on TNorm Lukasiewicz:

$$\text{Lower approximation} = (R_{B\ A}(y) = \text{INF}_{x \in x}\, \beta((R_B(x, y), A(x)) \quad (2)$$

Positive region based on lower approximation:

$$POS_B(y) = \left(U_{x \in x\, R_B R_d^x}\right) \quad (3)$$

Which are the sum lower approximations from similarity matrix for every topic. After that using positive region to calculating dependency function for each feature as in the Eq. 4:

$$\text{Dependency function} = \frac{\text{POSB(y)}}{\text{No. of messages}} \quad (4)$$

After obtaining the dependency value for each topic if then rule sets- based on resulting reducts from features selection will be building (Wang *et al*., 2007). In first iteration in this process will be creating empty set of rules and then in each time will add objects from Reducts set that have higher accuracy rule, this process will continue until one or both the following condition satisfying:

- Reached to maximum evolution function (or to degree α)
- Or when rules which its accuracy will not optimize classifier precision

Evaluation and experimental results: Confusion Matrix (CM) as one of the popular evaluation methods has been utilized for assisting the results of our research. The suggested system has been performed on 421 short messages with 70% for training purposes and 30% for testing purposes. The proposed method was compared against the popular algorithms the decision tree and Naive Bayes algorithms in addition to the traditional quick rules algorithm as shown in Table 4 and Fig. 2. The four confusion matrices in Table 5 and 6 have shown the return true positive decisions for the traditional quick rules and the enhancement after using the textual semantic analysis acted and semantic lexicon features. The results have shown encouraging enhancement in quick rules performance in particular when the semantic features have been applied with system. They have improved the topics quality that resulted from LSA as features selection and consequently improved the classifier performance.

Table 5: The confusion matrix for the quick rules with latent semantic analysis

	Predicted class	
Training 70% with synonyms	Not junk	Junk
Actual class		
Not junk	369	9
Junk	7	36

Table 6: The confusion matrix for the quick rules with latent semantic analysis and semantic lexicon

	Predicted class	
Training 70% with synonyms	Not junk	Junk
Actual class		
Not junk	371	-
Junk	15	34

CONCLUSION

Using semantic lexicon for enhancing the quick rules based fuzzy rough set classifier on short text messages was investigated in this research. The messages textual features are augmented with their semantic terms to improve the topics quality, LSA performance and the decisions making of recognizing the income messages correspondingly. The research has shown that aggregating the text messages with semantic lexicon has a significant positive impact on the classification results and the efficiency of the classification process. The accuracy and all others performance measures of the suggested classifier model using synonyms are better than that model without using it. Finally, this research has found that the accuracy rate is raised to 96.1% instead of 92.3% for the standard features only.

REFERENCES

Agrawal, P.K. and A.J. Agrawal, 2014. Opinion analysis using domain ontology for implementing natural language based feedback system. Intl. J. Inf. Technol. Comput. Sci., 2014: 61-69.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM., 55: 77-84.

Dave, K., S. Lawrence and D.M. Pennock, 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proceedings of the 12th International Conference on World Wide Web, May 20-24, 2003, ACM, Budapest, Hungary, pp: 1-10.

Deng, W.W. and H. Peng, 2006. Research on a naive bayesian based short message filtering system. Proceedings of the International Conference on Machine Learning and Cybernetics, August 12-16, 2006, Dalian, China, pp: 1233-1237.

Ding, Y., G. Zhu, C. Cui, J. Zhou and L. Tao, 2011. A parallel implementation of singular value decomposition based on map-reduce and PARPACK. Proceedings of the 2011 International Conference on Computer Science and Network Technology (ICCSNT) Vol. 2, December 24-26, 2011, IEEE, Harbin, China, ISBN:978-1-4577-1586-0, pp: 739-741.

Fodeh, S., B. Punch and P.N. Tan, 2011. On ontology-driven document clustering using core semantic features. Knowl. Inf. Syst., 28: 395-421.

Gomez, H.J.M., G.C. Bringas, E.P. Sanz and F.C. Garcia, 2006. Content based SMS spam filtering. Proceedings of the 2006 ACM Symposium on Document Engineering, October 10-13, 2006, ACM, Amsterdam, Netherlands, ISBN:1-59593-515-0, pp: 107-114.

Guan, P., W. Yuefen, C. Bikun and F. Zhu, 2012. K-means document clustering based on latent dirichlet allocation. Master Thesis, School of Economics & Management, Nanjing University of Science and Technology, Nanjing, China.

Holzinger, A., J. Schantl, M. Schroettner, C. Seifert and K. Verspoor, 2014. Biomedical Text Mining: State-of-the-Art Open Problems and Future Challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Andreas, H. and I. Jurisica (Eds.). Springer, Berlin, Germany, ISBN:978-3-662-43967-8, pp: 271-300.

Homayouni, R., K. Heinrich, L. Wei and M.W. Berry, 2004. Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinf., 21: 104-115.

Meng, L., R. Huang and J. Gu, 2013. A review of semantic similarity measures in wordnet. Intl. J. Hybrid Inf. Technol., 6: 1-12.

Miller, G. and W. Charles, 1991. Contextual correlates of semantic similarity. Lang. Cogn. Proc., 6: 1-28.

Price, R.J. and A.E. Zukas, 2005. Application of Latent Semantic Indexing to Processing of Noisy Text. In: Intelligence and Security Informatics, Kantor, P., G. Muresan, F. Roberts, D. Zeng and F.Y. Wang et al (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-25999-2, pp: 602-603.

Rada, R., H. Mili, E. Bicknell and M. Blettner, 1989. Development and application of a metric on semantic nets. IEEE Trans. Syst. Man, Cybern., 19: 17-30.

Sheu, P.C.Y., S. Wang, Q. Wang, K. Hao and R. Paul, 2009. Semantic computing, cloud computing and semantic search engine. Proceedings of the IEEE International Conference on Semantic Computing (ICSC'09), September 14-16, 2009, IEEE, Berkeley, California, USA., ISBN:978-1-4244-4962-0, pp: 654-657.

Sohn, D.N., J.T. Lee and H.C. Rim, 2009. The contribution of stylistic information to content-based mobile spam filtering. Proceedings of the ACL-IJCNLP 2009 International Conference on Short Papers, August 04, 2009, Association for Computational Linguistics, Stroudsburg, Pennsylvania, pp: 321-324.

Wang, X., E.C. Tsang, S. Zhao, D. Chen and D.S. Yeung, 2007. Learning fuzzy rules from fuzzy samples based on rough set technique. Inf. Sci., 177: 4493-4514.

Yin, Z., L. Cao, J. Han, C. Zhai and T. Huang, 2011. Geographical topic discovery and comparison. Proceedings of the 20th International Conference on World Wide Web, March 28-April 1, 2011, ACM, New York, USA., ISBN:978-1-4503-0632-4, pp: 247-256.