ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.12

Abbreviated Sentence Classification using Semi- Supervised Learning

DipikshaChourasiya[1], Ashish Tiwari[2]
M.Tech Student[1], Assistant Professor[2]
Department of CSE
VITS, Indore, India

Abstract:
Text classification is one of the major areas for researchers. Text mining is the special branch of data mining which deals with text and words. Digital communication mostly based on other small words those are not proper English. Sentences those have such type of words are difficult to be classified properly. This paper works on the classification of abbreviated sentences means sentences those have irregular words of English. For this purpose semi-supervised technique is used.

Keywords: text mining, text classification, semi-supervised learning.

I. INTRODUCTION

Area of data mining is related to analysis of data and collecting knowledge from that huge amount but data has several new spans and one of them is text. Text mining is the one field which focuses on text, word and sentences. Textual content analysis is todays need because conversations using digital media are enhanced. Text classification is the technique of text mining which selects and stores the particular text or word of their relevant class. Text classification requires some predefined classes in which it can put related words. This paper introduces the classification of abbreviated sentences. Today, use of online chatting and mobile text are increased and such type of conversation now is based on small and friendly words instead of actual English words. This paper introduces a semi-supervised learning technique for enhancing the accuracy of sentence classification. Semi supervised learning is a type of learning which uses both supervised and unsupervised leaning. In this system some of the known classes are introduced and some of the words are identified dynamically using online sites. This makes the system more efficient than traditional once.

II. PROBLEM DEFINITION

Text classification is one of the major issues in text mining area because it needs some extra efforts like human. Human can understand and respond properly about any word but that task should be played by text mining system that is the problem. In text communication human can understand the word meaning and abbreviation but text classification needs previous knowledge about that word.

Following problems may be raised during abbreviated sentence classification

- Sentence classification requires exact English sentences, but text communication in media uses the abbreviated sentences or user-defined short forms.
- Sentence classification requires word meanings by which that can be properly classified.
- Dynamic knowledge is required to learn new words.
- Domain based sentence classification is a tough task because one word may have different meanings.

III. LITERATURE SURVEY

Data mining has a number of branches according to their different aspects like spatial mining, web mining, sequence mining, text mining and so on. Text mining also has different terms like sentence mining, documents clustering or classification and word related operations. Many researchers are working in data mining domains. Some of the researchers embed the semi supervised way with the hierarchical clustering[1]. For the increase in accuracy and performance of present text clustering technique like hierarchical clustering, some predefined knowledge is added to clustering methods. In the document with abbreviated words, the domain name identification of document is also done by semi supervised learning based model [2]. In the area of text clustering, text contents need a meaningful explanation for the machine. This word explanation can be fetched from different websites [3]. Text mining is useful for document-briefing, this provides the exact meaning in less words for long textual contents [6] [7]. Text mining is the process of useful text retrieval from huge contents. Text mining is also useful for machine- human interaction systems [8]. Presently we are progressing towards the machine intelligence, so for understanding human language, text mining is the one way for this. Many researchers are trying to analyze the exact conversation through social media and microblogs on websites, but short forms of real English words are the main reason of decrement inpurity of accuracy [9]. Topic detection of text document and title extraction from huge textual contents are also the well-known area for data scientists [10]. Text clustering of microblog posts is also the area of research for the various text data analyzer. Text iteration model is proposed as a solution by some researchers [11]. Social media blogs can be analyzed but some of the duplicate blogs can also be identified by using text analysis techniques [12]. Text mining approaches in various fields are the presented by milos in their paper. They describes the various applications where text mining will be helpful [14].

IV. METHODOLOGY

Sentence classification is the variation of text classification because sentence is the group of texts. This paper introduces the method for classifying sentences those includes abbreviated words. For classification of such type of sentences, firstly we should have knowledge about abbreviated words which are used in that sentence. In chatting we generally use short forms of actual word, those are not properly from any language and these are only used for simplification of text conversation. For identification of those words correctly this paper uses the method which contains online web assistance for correct classification of sentences. This paper proposes the semi supervised method for achieving quality classification.

Semi-Supervised Learning: Supervised learning means all the directions are given there and unsupervised learning means system learns by its own efforts, but this paper uses semi supervised learning for achieving goals of abbreviated sentence classification. Semi-supervised learning is the learning which includes some of the defined instructions and some of the own knowledge.

Classification with semi-supervised learning: This paper shows a method for classification of abbreviated sentences. Classification is the process of arranging objects into given groups. But in the case of sentence classification there should be exist some knowledge of word definition or meaning, those are used in that sentence. Classification means already classes are defined there, but we are using semi supervised strategy for learning the system means some of the knowledge are stored in the database for classification help but meaning and definitions of all words are not available in the database so this system also uses the online dictionary help for this. Remaining will also be done by using all their combined knowledge.

The proposed system may work on following steps:

Document: Initially, a document is used as input that contains group of abbreviated sentences. This document may contain chatting logs or other small conversations.

Conversion: This is the second step in which abbreviated sentence is converted into its actual form. There are two ways for doing conversion of abbreviated form to actual form. First way, by using predefined database and second one by using online dictionaries or acronym websites, this is the dynamic approach.

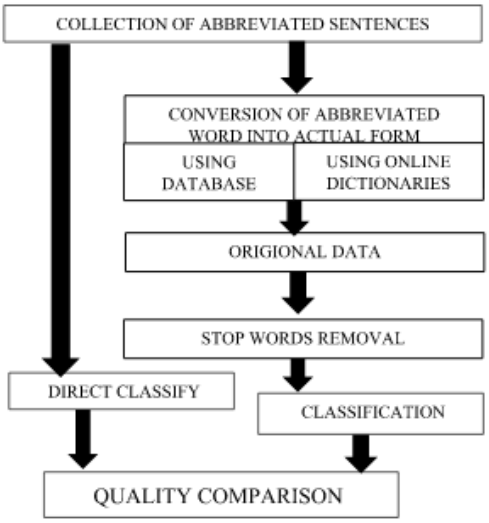


Figure.1. Proposed model for improved clustering

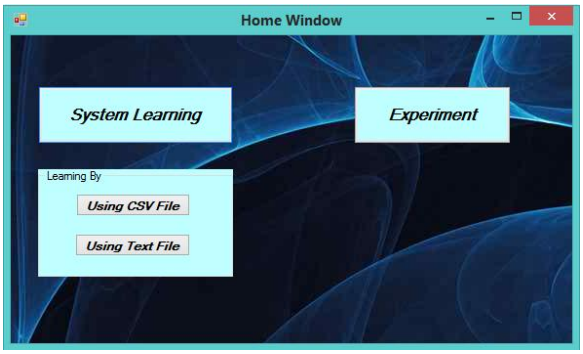
Original Data: After conversion from abbreviated form to its approx. actual form by using existing database or using websites, the document has original data.

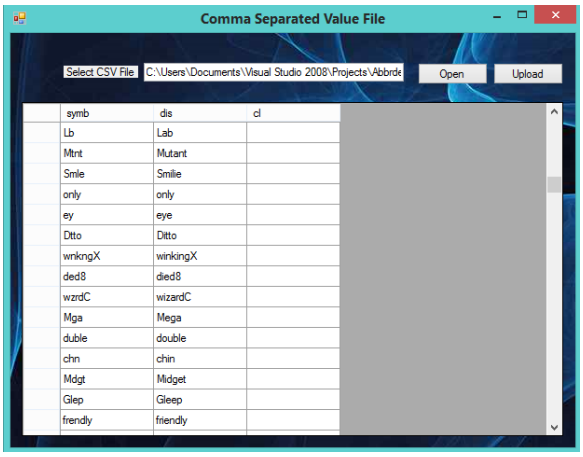
Removal of Stop-words: Original data also includes some unwanted words which are not required for analysis purposes.

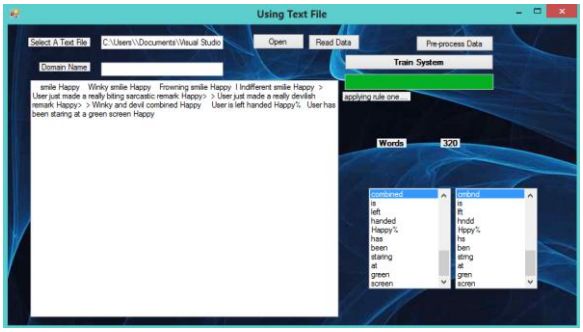
Comparison: Last step of proposed method is to compute the performance of method by using some distinct parameters and compare them with different traditional methods.

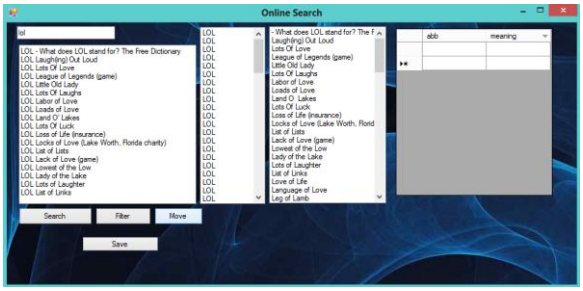
V. IMPLEMENTATION

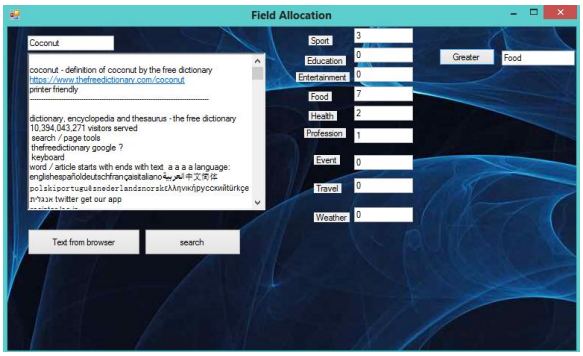
The proposed system is developed using dot net technology. First form of GUI is the login page for providing security to the software. Next form is the home form that has different options out of which user can choose anyone.

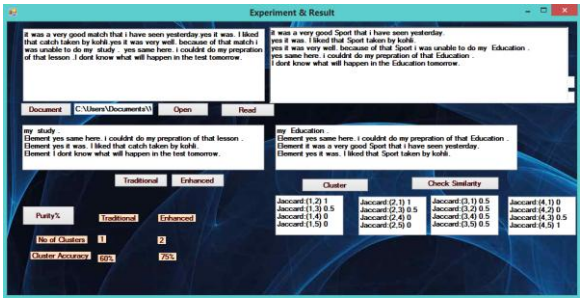




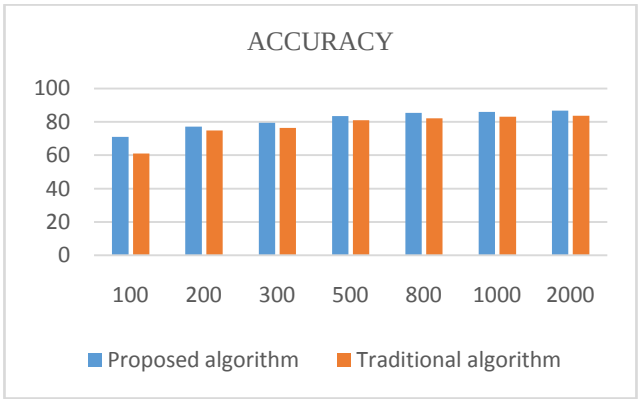








VI. RESULTS



VII. CONCLUSION

This paper introduced a new sentence classification technique. This technique uses the semi supervised learning for preparing system knowledgebase. Semi supervised way of classification is fruitful for grouping of abbreviated word. This work enhances the purity of sentence classification in the case of abbreviated sentences.

VIII. REFERENCES

[1].LiZhengand Tao Li. (2011). *"*Semi-supervised Hierarchical Clustering,*"* 11th IEEE International Conference on Data Mining.

[2]. Sachin Malviya, Dr. Pramod S. Nair, "A new approach of semi-supervised clustering with abbreviation detection and domain prediction using online dictionaries", IJESC Volume 6, Issue 7.

[3]. Huma Khan, Mayur Rathi, "Short Text Clustering Using Web Content Mining: A Survey", IJESC Volume 7, Issue 7.

[4]. Vishal Gupta, Gurpreet S. Lehal. (August 2009), "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1.

[5]. P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar (July 2008) "Knowledge Extraction Using Rule Based Decision Tree Approach", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7.

[6]. Andreas Hotho, Andreas Nurnberger, Gerhard Paab, Fraunhofer AiS (May 13, 2011) "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin.

[7]. Hien Nguyen, Eugene Santos, and Jacob Russell (November 2011) "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 41, No. 6.

[8]. Umajancy. S, Dr.Antony Selvadoss Thanamani (August 2013) "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8.

[9]. Zintao Lio and Wei Chen "Short Text Feature Selection for Micro-blog Mining" IEEE, 2010.

[10]. Siqi Huang, Yitao Yang, HuakangLi, GuoziSun, "Topic Detection from Microblog Based on Text Clustering and Topic Model Analysis" IEEE 2014.

[11]. Kai Gao, Bao-quan Zhang "Modelling on Clustering Algorithm Based on Iteration Feature Selection for Micro-blog Posts" ICMIC (2014).

[12]. Qunyan Zhang, Haixin Ma, Weining Qian, Aoying Zhou "Duplicate Detection for Identifying Social Spam in Microblogs" IEEE 2013.

[13]. Yihong Rong, Jia Song, "Mining A Government Affairs Microblog Network on Sina Weibo with Social Network Analysis", IEEE-2013.

[14]. Miloš Radovanović, Mirjana Ivanović (2008). "Text Mining: Approaches and Applications", Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 227-234.