A System to Filter Undesired post From OSN user Space using NLP

Sayana S [1*], Soorya P [2]

[1*]PG Scholar, Dept. of computer science and engineering Professor, Dept. of computer science and engineering TKM Institute of Technology
Kollam,India

Abstract: **One of the fundamental issue faced in the current work is people posting the indecent messages in private wall of his/her contacts which annoys other people to seeing them. In order to avoid this an automated system called filtered wall is used to filter the unwanted message from OSN user space. To achieve this, a flexible rule based system and a short text classifier is introduced. This rule based system is based on machine learning techniques. Machine learning technique learns from training data and use these training data to create classifier for classifying the new data. For this purpose Bayesian algorithm is used. Bayesian classifier is considered as one of the known algorithm for text classification. In this work a new feature introduced called semantic interference using NLP. This system function as a warning system. One goal of NLP is to measure the quality of the system in order to determine whether the system answers the goal of the user.**

Keywords-Online social network, short text classifier, Machine learning, Semantic similarity.

I. INTRODUCTION

One of the best entertainment of the younger generation is in the form of online social network sites. Online social network is a platform to build social network or social relation among different peoples in a network in order to communicate and share different types of contents like text, images, audios, videos etc. The previous work presented an automated system called filtered wall (FW). This filtered wall helps to prevent the indecent words from the OSN user space. This is achieved through machine learning text categorization technique. Machine learning text categorization technique which automatically assign each text messages to a set of categories based on its contents.

Previous work explained the way of filtering and necessity of filtering in the OSN user wall. But does not discussed about the meaning of the word in a post. Classification technique in data mining consist of different concepts extracting from training data set in order to differentiate one class from other classes. One of the main goal of classification technique is to provide better result in terms of accuracy and precision. In many cases classification technique cannot provide better accuracy due to huge dataset. This paper concentrated on splitting the training data and then applies Bayesian classification with each group in order to get better result in terms of precision and recall.

II. RELATED WORKS

A number of methods have been developed for classifying objects to different class by using a learning classifier. Ana Fernandez et al[2] mentioned about the strength of relationship between two friends in facebook. They proposed a model to infer context from social network by the application of Natural language processing techniques(NLP) and data mining techniques. Jiawei et al[7] discussed about novel class detection problem of a classifier. Most of the classification technique that ignore one important aspects called arrival of novel class detection. They provide a solution called multi class framework for the novel class detection problem and distinguish the new or novel class with existing classes. John carol et al[6] developed a cross domain sentiment classifier using a sentiment sensitive thesaurus. Sentiments are expressed differently in different domains. So there is a need of different classifier for different domains. In order to avoid this mismatch problem cross domain classifier is introduced. Cross domain classifier is performed well in different domains by using sentiment sensitive thesaurus. Ronald adam et al[4] introduced a semantic interference using NLP in order to find out the words which have similar meaning. Gabar et al[3] mentioned about the implementation of soft computing techniques in NLP. They tried to implement a part of speech tagger to extract and categorization of words based on their meaning and functions. Dr Alaan et al [1] mentioned about semantic similarity between words based on wordnet semantic dictionary and also introduced some algorithms to capture the semantic relation between the words in a text. M. Zubair Shafiq et.al [10] discussed about leaders and followers in online social networks. Identifying leaders and followers in online social networks is important for various applications in many domains such as advertisement, community health campaigns, administrative science, and even politics. M. Vanetti et.al [8] proposes a content based filtering system. The explosive growth of online information demands new technique for prioritizing and presenting items of potential interest to users.

III. SYSTEM ARCHITECTURE DESIGN

In this work, a new feature called semantic interference using NLP is introduced. Semantic similarity measure is a central issue in artificial intelligence. It is widely used in natural language processing(NLP). Semantic similarity play an important role to capture semantic relation between two different words and thereby produce more reliable result.. To decide the semantic similarity between two different words for a computer, it should understand the semantics of the words.

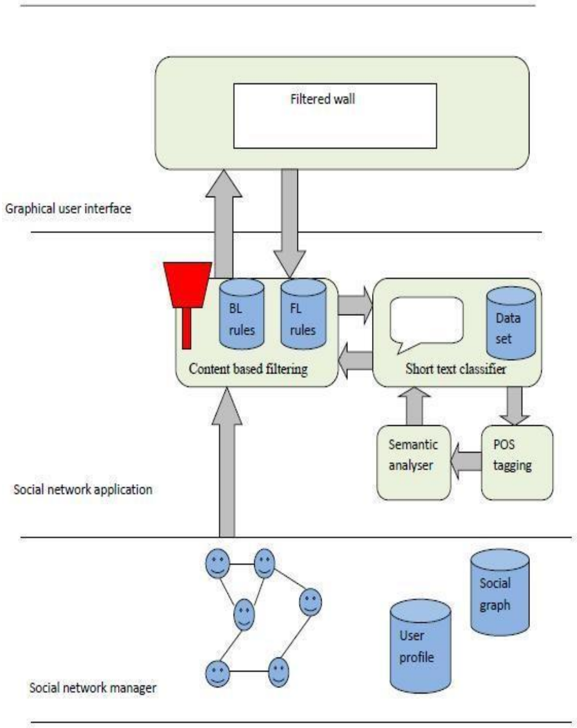

Fig 1:Architecture of OSN based on NLP

IV . IMPLEMENTATION

NLP includes the consideration of semantic knowledge. There are several semantic theories exist.

i Natural language processing

Natural language processing is an efficient algorithm to process text and to make the information accessible to computer applications. It uses a collection of technique to extract grammatical information and meaning from input in order to perform useful text. It is an automatic approach to process human natural language. Most of the NLP applications contains POS tagger as one of its core component. There are many advantages of natural language processing as a communication channel between man and machine. One of them is that the man already knows the natural language, so that he does not have to learn an artificial language nor bear the burden of remembering its conventions over periods of disuse. There arise occasions where he knows what he wants the machine to do and can express it in natural language, but does not know exactly how to express it to the machine. In order to overcome those a new facility introduced called natural language understanding. A facility for machine understanding of natural language could greatly facilitate the efficiency of expression both in speed and convenience.

ii Natural language understanding

Natural language understanding consist of four section.

1. Morphology

It is the first stage of analysis once input has been received. It looks at the ways in which words break down into their components and how that affects their grammatical status.

2. Syntax

It involves applying the rules of the target language's grammar, its task is to determine the role of each word in a sentence and organize this data into a structure that is more easily manipulated for further analysis.

3. Semantics

It is the examination of the meaning of words and sentences. Semantics convey useful information relevant to the scenario as a whole.

4. Pragmatics

They are the sequence of steps taken that expose the overall purpose of the statement being analyzed. This will be broken down into Ambiguity and Disambiguation to facilitate understanding.

iii POS tagger

A Part-Of-Speech Tagger is a piece of software that reads text in some language and assigns parts of speech to each word such as noun, verb, adjective, etc., POS tagger plays an important role for most of the NLP applications. POS tagging is the process of classification of words according to their meaning and function. Part of speech is the process of marking up a word in a text as corresponding to a particular part of speech based on both its definition, as well as its context.

iv Semantic analyzer

Computers are very fast and powerful machines. However, the computer which process all kinds of text written by human beings. One of the main goal of language analysis is to obtain a suitable representation of text structure and thus make it possible to process text based on their content. Semantic analysis includes consideration of semantic knowledge. Once the computer has arrived at an analysis of the input sentence's structure, a semantic analysis is needed to ascertain the meaning of the sentence. Mapping a sentence into logical form or meaning is called semantic interpretation.

Sometimes some semantic ambiguity may arise. When more than one possible meaning exists for a sentence as in "He lifted the branch with the red leaf." It may mean that the person in question used a red leaf to lift the branch or that he lifted a branch that had a red leaf on it.

After determining the structure of sentence the next step is to determine the meaning of each word in a sentence. This semantic process build up a representation of each word in a sentence and then describe the action performed by each word in a sentence. And also gather information in order to determine which meaning that are intended by the user.

v Disambiguation

There are many techniques and tools to decide which interpretation of a word to use, some of these techniques are listed below:

1.Prior probabilities

These are rules that tell the system that a certain word phrase nearly always means a certain thing without looking at anything else, this is a purely statistical approach to disambiguation.

2.Conditional probability

It examines the scenario in reference to the origin of the phrase in order to make the decision on the meaning of a word phrase.

3.Context

It looks at the environment and incidents surrounding the phrase in order to make a decision on which interpretation to use.

4.World Models

They are needed for a good disambiguation system, to allow for the selection of the most practical meaning of a given sentence. This world model needs to be as broad as the scenarios the system would encounter in its normal operation.

V. .PERFORMANCE EVALUATION

Aim of this work to enhance the flexibility of the system by adding the feature of semantic similarity using NLP. Previous work only focused on content based filtering not focuses on semantic similarity between the words. In order to find semantic similarity, different types of semantic similarity technique are used. To interpret natural language sentence correctly, the system will not only have to parse the sentence but also associate the words with things in the text or word.

VI. RESULT

This system act as a warning system. If any user trying to send any vulgar message, then the proposed system will show an alert like a text message. System based on automatically learning the rules that can be made more accurate than by simply supplying the input data. It successfully extract meaning and relationship between different words by using semantic similarity measure based on NLP. Natural Language systems hold great promise in areas involving human computer interaction.

VII. CONCLUSION

Measuring semantic similarity in textual items are considered as the core part of natural language processing applications. One of the advantage of this semantic similarity measure is that it operates at different levels eg: a single word or entire document. Previous work which filter the unwanted contents on the OSN user wall by using content based filtering but not based on semantic relatedness between the words. Here semantic similarity measure based NLP is used to find the similarity between words.

REFERENCES

[1] Towards Exploring Semantic Similarity based on WordNet Semantic Dictionary International Journal of Computer Applications March 2013.

[2] Inferring Contexts From Facebook Interactions: A Social Publicity Scenario Sandra Servia-Rodríguez, Ana Fernández- Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Arias ieee transactions on multimedia, vol. 15, no. 6, october 2013.

[3] Natural Language Processing based Soft Computing Techniques Jabar H. Yousif Faculty of Computing and Information Technology Sohar University, P.O.Box.44, P.C.311, Sohar, Sultanate of Oman International Journal of Computer Applications (0975 – 8887) Volume 77 – No.8, September 2013.

[4] plagiarism detection algorithm using natural language processing based on grammar analyzing angry ronald adam,suharjito journal of theoretical and applied information technology 10th may 2014. vol. 63 no.1

[5] Measuring Semantic Similarity between Words Using eb Pages T.Sujatha, Ramesh Naidu G, P.Suresh B International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012. [6] Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus Danushka Bollegala, David Weir, and John Carroll ieee transactions on knowledge and data engineering, vol. 25, no. 8, august 2013. [7] Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints Mohammad M. Masud, and Bhavani Thuraisingham, Fellow, ieee transactions on knowledge and data engineering [8]Content-based Filtering in On-line Social Networks M. Vanetti, E. Binaghi, B. Carminati, M. Carullo and E. Ferrari ieee transactions on knowledge and data engineering. [9] Fully Anonymous Profile Matching in Mobile Social Networks Xiaohui Liang, Xu Li, Kuan Zhang, Rongxing Lu, Xiaodong Lin, and Xuemin (Sherman) Shen, ieee journal selected areas in communications/supplement. Identifying Leaders and Followers in Online Social Networks M. Zubair Shafiq, Student Member, IEEE, Muhammad. [10] Identifying Leaders and Followers in Online Social Networks M. Zubair Shafiq, Student Member, IEEE, Muhammad U. Ilyas, Member, IEEE, Alex X. Linu journal on selected areas in communications/supplement, vol. 31, no. 9, september 2013.