UNIFIED EXTENSIBLE FRAMEWORK FOR SENTIMENT ANALYSIS BASED ON OPINION DISAMBIGUATION

Aarthi K[1], Bhavadharani T[2], Deepika J[3] ,Karthika R[4](UG Students),

Lakshmi R[5](Assistant Professor)

B.Tech ,Department of Information Technology ,SRM Valliammai Engineering College, Kattankulathur,Kanchipuram,Tamil Nadu, India.

ABSTRACT

Sentiment Disambiguation is a vital component of any sentiment analysis or opinion analysis of the Data Mining domain. It plays a decisive feature to distinguish sentiment polarity of different text granularity. Another important point is to process the fine-grained large data sentiment analysis based on sentiment disambiguation lexicon to improve the accuracy of sentiment prediction. This system uses the Statistical Natural Language Processing computational techniques for the sentiment analysis. The existing system is based on a customized fuzzy method approach with two-stage training to deal with text ambiguity. The existing system classifies three types of cyberhate speeches: race, disability, and sexual orientation. In the existing system, the prediction accuracy has decreased when large data set contains huge diversity in the text. The existing system cannot achieve the efficiency of removing the ambiguity when there is intersectionality among a wide variety of cyberhate speech. The proposed system implements the process to determine the equivocal text into various categories toward Sentimental Disambiguation which narrows down the meaning of the words. The proposed system combines the sentiment label data and sentiment contrast data between focus keywords and context words into words vector representation learning model. The accuracy of the proposed prediction model will be very high even though the large diversity in the input text has fuzziness, ambiguity, and vagueness.

Keyword : - *Sentimental Analysis, Text Tokenization, Cyberhate, Text Summarization.*

1.INTRODUCTION

Sentiment analysis can be used refer to many different but related problems. Emotion measurements are a challenging and debated task. Sentiment analysis also called opinion mining or polarity detection. It is also used to refer to the task of automatically determining the valence of a piece of text. SA is involved in determining one's attitude towards a particular target or topic. The majority of these valence and emotion classification approaches employ statistical machine learning techniques, although some rule-based approaches also persist. Automatic detection and analysis of affectual categories in the text have wide-ranging applications. Automatic methods to detect various affect categories become more accurate, their use in natural language applications will likely become even more ubiquitous.

Steps of sentiment analysis:

1)Initialization step (data collection, data processing, attribute selection)

2)Learning step (algorithm, training model) and

3)Evaluation step (test set).

With the rise of Web 2.0, microblogging websites have increasingly become a valuable platform for people to express their opinion and sentiment on a certain topic. Blogs, forums and social media platforms allow users to easily add blog posts, reviews, reactions and ratings to share their point of view on the internet.

2.LITERATURE SURVEY

2.1. Guo-Shuai Liu, Rui-Qi Wang, Fei Yin, Jean-Marc Ogier, Cheng-Lin Liu, "Fast Genre Classification of Web Images Using Global and Local Features" introduced genre classification based on the web images available in the web. Using two stages by defining global features and local features using SVM. Two fusion strategies is used by train the second stage classifier and generate the final prediction depending on the usage of local features in the second stage. To validate the effectiveness of our proposed method, we built a database containing more than 55,000 images from various sources. On our test image set, we obtained an overall classification accuracy of 98.4% and the processing speed is over 27FPS on an Intel(R) Xeon(R) CPU (2.90GHz).

2.2. Tu manshu and Zhao Xuemin researched cross-domain sentiment classification, "**An End to End Model for Cross-Domain Sentiment Classification"**(CDSC) aims to adopt a model trained by a source domain to a target domain. It has received considerable attention in recent years. Most existing models mainly focus on learning representations that are domain-independent in both the source domain and the target domain. However, domain specific features, which should also be informative are ignored by these models. This paper proposes an end to end model. It can capture both the source domain and target domain features at the same time. This model includes two parts; one is a cloze task network (CTN), we use it as an auxiliary task to fine-tune words embedding in both domains. Another is a Convolutional hierarchical attention network(CHAN), we use it for sentiment classification. The CHAN can capture important words and sentences concerning sentiment based on its two stages of attention mechanism. The CTN and CHAN conduct jointly learning we abbreviate this model as CCHAN. The experiments on the Amazon review datasets demonstrate that the proposed CCHAN can significantly outperform the state-of-the-art methods.

2.3. Yuan Chi, Elias J. Griffith, John Yannis Goulermas ,"Binary Data Embedding Framework for Multiclass Classification "have researched a novel manifold embedding method for the automated processing of large varied datasets. The method is based on binary classification, where the embeddings are constructed to determine one or more unique features for each class individually from a given dataset. The proposed method is applied to examples of multiclass classification that are relevant for large-scale data processing for surveillance (e.g., face recognition), where the aim is to augment decision making by reducing extremely large sets of data to a manageable level before displaying the selected subset of data to a human operator. The method consists of two stages: Preprocessing and embedding computation. In the embedding computation, adaptive measures of intraclass and interclass information are proposed, based on the concepts of "friend closeness" and "enemy dispersion". Besides an indicator for the weighted pairwise constraint is proposed to balance the contributions from different classes to the final optimization,to better control the relative positions between the important data samples from either the same class (intraclass) or different classes (interclass). The effectiveness of the proposed method is evaluated through comparison with seven existing techniques for embedding learning, using four established databases of faces, consisting of various poses, lighting conditions, and facial expressions, as well as two standard text datasets.

3. EXISTING SYSTEM

The existing system is based on BOW. BOW is one of the most popular methods of feature extraction, it has a few limitations that could affect the performance of learning from textual instances. In particular, from semantic perspectives, the same word may have different meanings, which could lead to the case that a word could be highly relevant to the positive class in some cases but also highly relevant to the negative class in other cases. For example, the word "deserve"can be used to praise students who work hard by saying "You fully deserve the success," whereas the same word can be used to criticize students who failed due to low motivation by saying "That is what you deserve." Also, from syntactic perspectives, the same word may act as different parts of speech. For example, the word "approach" could be both a verb and a noun, which could lead to different abilities to discriminate between classes. In particular, when the above word is used as a verb, it could lead to a negative message such as "I approach you to do something for me."

In contrast, when the word is used as a noun, it would generally show a neutral meaning. The above-mentioned two points indicate that when a word has different meanings or acts as different parts of speech, it is not appropriate to simply treat the word as a single feature. Due to the limitations of BOW, researchers have been motivated to use, which is aimed at combining n sequential words as a feature instead of a single word and has led to

the enrichment of semantic information with improvements of classification performance. In this context, the value of each feature is also represented by different types of frequency, such as corpus frequency (CF), DF, and sentence frequency (SF), apart from the commonly used ones (TF, IDF, and TF-IDF). In particular, CF represents the frequency of an n-gram in the whole corpus, whereas DF/SF represents the number of documents/sentences in which an n-gram appears.

3.1 DISADVANTAGES

- Difficult to determine creative portion of languages such as sarcasm, irony, humor and metaphor.
- Troublesome to interpret creatively spelled social media words, emoticons, abbreviations, hashtagged words, etc.
- Lack of usage of large amount of training set data.
- Increased time complexity and space complexity.

4. PROPOSED SYSTEM

Semantic Text Similarity plays a major role in natural language processing. In recent years, researchers have paid considerable attention to Semantic Text Similarity. Some breakthroughs have been made in English, but there are few disadvantages when these models are applied to English: Single sequence models don't consider semantic ambiguity such as polysemy, synonym.These models don't consider that English stop words are important for English word segmentation, voice analysis, semantic understanding. Firstly, in order to overcome the first problem, we proposed the double short text sequences model that has two identical processing two text sequences at the same time.

The proposed query system analyses the sentiment of citizen by sentiment analysis. According to that, priorities to the given citizens on the basis of the intensity of the citizen complaint. Understanding the short text is the major challenge in the system like short texts do not follow the syntax of written language, short text does not have sufficient statistics to support for approaches like text mining, short text is ambiguous and noisy.

In this system to understand natural language semantic knowledge is provided by the knowledgebase. This system will help many organizations to ensure quality service provision and customer satisfaction with less human efforts.

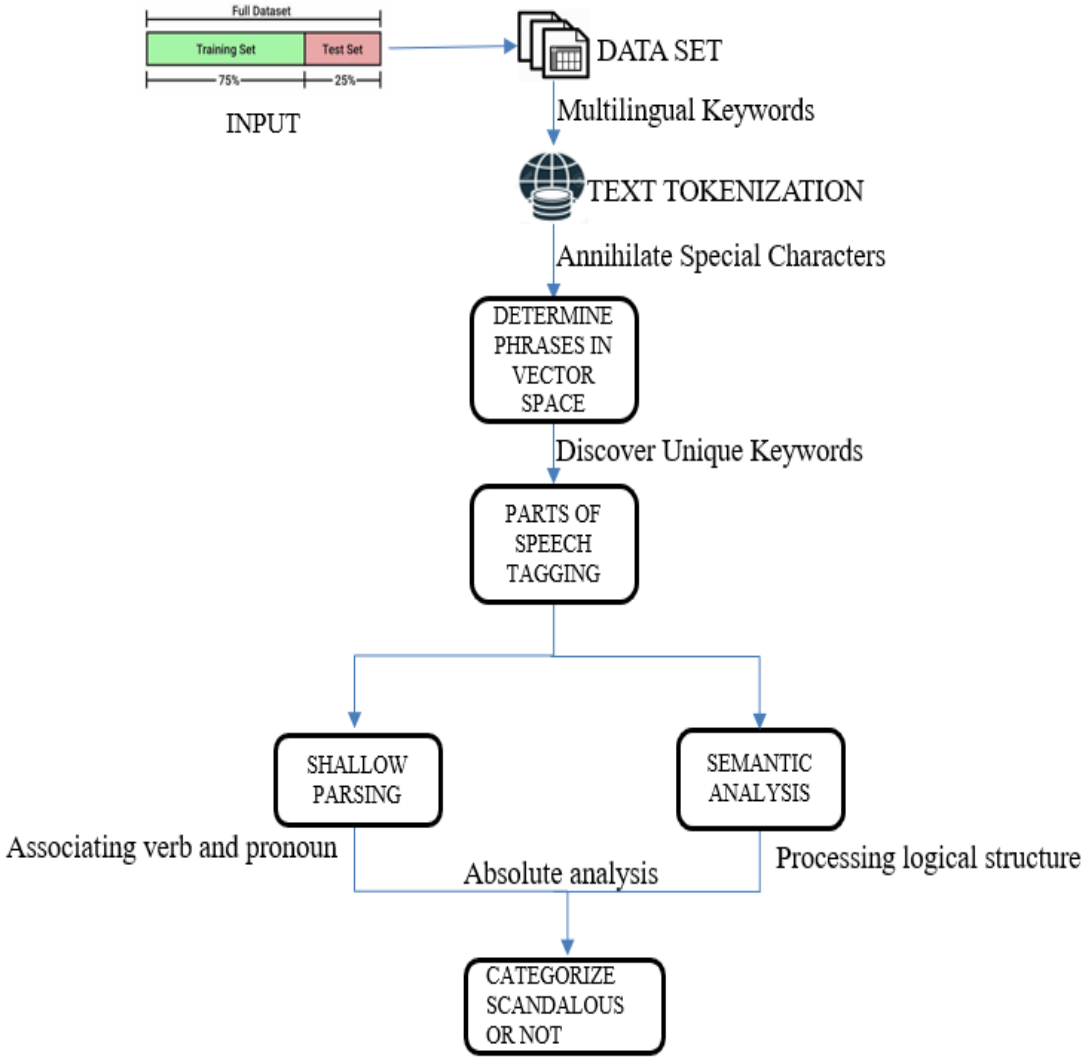
4.1.ADVANTAGES

- Accuracy of determining the emotion is more.
- Analyze the overall semantics of a tweet.
- Text summarization is the key for efficiency.

5. IMPLEMENTATION

Initially, the enormous amount of finite dataset is given as input. Then , text tokenization process is carried out to split the sequence of strings into parts of strings such as words, phrases, symbols like emoticons, special characters, etc. Discovering unique root words and representing them through vector space models. This vector space model is used to rank the relevant documents for finding document similarity. Next, the syntactic and semantic properties of each tweets are analyzed. After analysing the tweets completely ,the sentiment of the corresponding tweets are envisioned.

6. BLOCK DIAGRAM



7. HARDWARE REQUIREMENTS

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15'6 Colour Monitor.
- Ram : 512 Mb.

8. SOFTWARE REQUIREMENTS

- Anaconda-Jupyter Notebook : Open source distribution of python that aims to simplify package deployment and management. An application used to create and edit documents that display input and output of python script.

- NLTK : A suite of libraries and programs for NLP in python.

9. CONCLUSION

The main contribution of this paper are summarized in the following three aspects: Firstly, a semantic text similarity model for double short English sequences is proposed, which overcomes the shortcomings that the single sequence model can't handle polysemy and synonym. Secondly, according to the English characteristics, the semantic similarity data set of English short texts is constructed and the Stop words are reserved during the model training. Thirdly, we test the replaceability and transitivity of the model respectively in the two test data sets. The results show that the model has a great improvement in transitivity test and a certain improvement in the replacement test, which shows that the Generalization ability of the model has been improved.

10.REFERENCES

- R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in Proc. 17th Int. Conf. Distrib. Comput. Netw., Jan. 2016, pp. 4–7.
- Y. Wang, X. Zheng, and X. Hu, ''Short text sentiment classification of high dimensional hybrid feature based on SVM,'' Comput. Technol. Develop., vol. 28, no. 2, pp. 88–93, 2018.
- H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on Arabic social media," in Proc. 1st Workshop Abusive Lang. Online, Vancouver, BC, Canada, Aug. 2017, pp. 52–56.
- Neviarouskaya and M. Aono, ''Sentiment word relations with affect, judgment, and appreciation,'' IEEE Trans. Affect. Comput., vol. 4, no. 4, pp. 425–438, Oct./Dec. 2014.
- Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," in Proc. IEEE Int. Conf. Fuzzy Syst., Naples, Italy, Jul. 2017, pp. 1–6.
- H. Lee and S. Kang, "Spam message filtering by using Sen2Vec and feedforward neural network", 4th Annual Conference on Computational Science & Computational Intelligence(CSCI 2017), pp.123-123, 2017.
- Y. An, S. Sun, and S. Wang, ''Naive Bayes classifiers for music emotion classification based on lyrics,'' in Proc. Int. Conf. Comput. Inf. Sci. (ICIS), Wuhan, China, 2017, pp. 635–638.
- R. Wang et al., ''Research of text sentiment classification based on improved semantic comprehension,'' Comput. Sci., vol. 44, no. 11A, pp. 92–97, 2017.
- S. Kang, "A normalization method of distroted Korean SMS sentences for spam message filtering," KIPS Transaction and Data Engineering, Vol.3, No.7, pp.271-276, 2014.
- H. Liu and M. Cocea, "Fuzzy rule based systems for interpretable sentiment analysis," in Proc. Int. Conf. Adv. Comput. Intell., Doha, Qatar, Feb. 2017, pp. 129–136.
- S. Kiritchenko, X. Zhu, and S. M. Mohammad, ''Sentiment analysis of short informal texts,'' J. Artif. Intell. Res., vol. 50, no. 1, pp. 723–762, 2014
- F. Huang et al., ''Mining topic sentiment in microblogging based on multifeature fusion,'' J. Comput., vol. 40, no. 4, pp. 872–888, 2017.