IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper | Vol.-8, Issue-1, Jan 2020 | E-ISSN: 2347-2693

Data Classification Approach For Text Analysis and Its Ambiguity

Supriya M. Yawalkar[1*], A. S. Kapse[2]

[1,2]Dept. of Computer Science and Engineering, P. R. Pote (Patil) College Of Engineering And Management, Amravati, Maharashtra, India

Corresponding Author: Supriyay66@gmail.com

DOI: https://doi.org/10.26438/ijcse/v8i1.141145 | Available online at: www.ijcseonline.org

Accepted: 20/Jan/2020, Published: 31/Jan/2020

***Abstract*:** Sentiment analysis or opinion mining is one of the fastest growing fields with its demand and potential benefits that is increasing every day. With the onset of the internet and modern technology, there has been a vigorous growth in the amount of data. Each individual is able to express his/her own ideas freely on social media. All of this data can be analysed and used in order to draw benefits and quality information. In this paper, the focus is on cyber-hate classification based on for public opinion or views, since the spread of hate speech using social media can have disruptive impacts on social sentiment analysis. In particular, here proposing a modified fuzzy approach with two stage training for dealing with text ambiguity and classifying three type approach positive, negative and neutral sentiment, and compare its performance with those popular methods as well as some existing fuzzy approaches.

Keywords: Ambiguity, cyber hate, fuzzy, Sentiment analysis.

I. INTRODUCTION

Sentiment analysis identifies the attitude or mood of people through natural language processing, text analysis, and computational linguistics. In recent years, machine learning has become a very powerful for classifying sentiments. In particular, support vector machines ,naïve bayes, decision trees, deep neural networks kind of method used to classify the text into categories and to detect cyberbullying detection [1], [2], abusive language detection [3], [4], movie reviews [5], [6], and cyberhate identification [7], [8]. In the context of machine learning, the above-mentioned algorithms are all aim to distinguish between one class and other classes. However, in the context of text classification, different classes are truly mutually exclusive, instances could be very complex and are thus difficult to be classified uniquely to only one category. This has motivated researchers to develop fuzzy methods for text classification, which are able to deal with fuzziness, imprecision, and uncertainty of text.

In this paper, focus is on the detection of online hate speech (cyberhate) in text posted to social media platforms on the basis of four types of online hate speech, namely, religion, race, disability, and sexual orientation, by proposing a novel fuzzy approach, especially for dealing with text ambiguity. The proposed fuzzy approach is different from existing fuzzy systems in two aspects. First, traditional fuzzy approaches typically aim at the production of single classifiers. In this aspect, the proposed fuzzy approach involves fusion (combining the membership degrees for each class) of multiple fuzzy classifiers produced with different parameters setting. Second, traditional fuzzy approaches generally employ a fixed rule to provide a distinct class label as an output. In contrast, proposed fuzzy approach involves a semifixed rule of defuzzification. In both of the two aspects, the proposed fuzzy approach can achieve effective disambiguation of text. Therefore, the bias of a single fuzzy classifier on the majority class (nonhate class) is much reduced, leading to reduction of the false negative rate.

Motivation of the research is because, although bag of word is one of the most popular methods of feature extraction, it has a few limitations. 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. This point indicates that when a word has different meanings, it is not appropriate to simply treat the word as a single feature. The proposed system overcomes this problem.

The rest of this paper is organized as follows. Section II describes related work that is relevant to cyberhate research and fuzzy classification. In Section III, presenting the proposed fuzzy approach and illustrate the procedure of fuzzy classification. In Section IV presents conclusion and next appears the references.

II. LITERATURE REVIEW

This section involves a review of feature extraction methods used for preprocessing of textual data, an overview of cyberhate research in the context of machine learning-based text classification, and the background of fuzzy text classification.

Prior work has shown sentiment analysis of social data can be used to predict movie revenues, Asur and Huberman [23] showed that analysis of sentiment content on urls, retweets and their hourly rates of Twitter can predict boxoffice movies revenues. Exploring cultural and linguistic differences in ratings and reviews [24], correlate with contemporaneous and subsequent stock returns [25], sentiment evolution in political deliberation on social media channels [26], assess sentiment towards a new vaccine and explore semantic-level precedence relationships between participants in a blog network [27]. To briefly expand, [28] proposed a methodology for the detection of bursts of activity at the semantic level using linguistic tagging, term filtering and term merging. They used a probabilistic approach to estimate temporal relationships between the blogs.

A. Review of Feature Extraction Methods-: In general, there are two popular methods that have been applied in feature extraction for sentiment analysis and cyberhate detection, namely, bag of words (BOW) and nGrams (NG). BOW extracts a bag of distinct words for textual data, and each of the words is used as a feature. Although BOW is one of the most popular methods of feature extraction, it has a few limitations. In particular, from semantic perspectives, the same word may have different meanings, 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 NGrams [12], [13], 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.

B. Overview of cyberhate Research-: Since cyberhate has been considered as a legal issue in many countries researchers have thus been motivated to develop tools for automatic detection of hate speech, in order to manage effectively posts containing hateful contents. Various methods of feature extraction and learning algorithms have been used for advancing this area. The naïve bayes algorithm was used in [13] for training classifiers on unigram (BOW) features, toward examining every single word for judging whether the tweets were fully hateful or not. In contrast, Mahmud et al. [14] focused their work on examining the sentence structures, which tend to be indicative of offensive remarks. In particular, convolutional neural networks (CNNs) have been used in [16] for classifying hate speech with different types of word vectors as features.

C. Background of Fuzzy Text Classification-: Fuzzy classification, which is based on fuzzy logic, is aimed at dealing with linguistic uncertainty that is involved in instances. Furthermore, the use of the fuzzy rule-based approach was investigated in [6] for multisentiment analysis and the results showed that the fuzzy approach could provide more refined outputs by reflecting different intensities of sentiment. In this context, each instance is typically not clear-cut and thus belongs to different classes. In recent years, fuzzy approach was proposed in [17] toward automatically building a corpus that can be used for comparison of text similarity. The experimental results showed that the fuzzy metrics had a higher correlation with human ratings in comparison with other traditional metrics. An unsupervised fuzzy approach was used in [18] toward achieving gender-based classification of Twitter users. An automatic approach based on a semantic similarity measure was introduced in [20] for recognizing emotion context from online social networks in the setting of fuzzy classification.

III. PROPOSED SYSTEM ARCHITECTURE

In proposed system there are two modules are available. First module indicate text analysis framework. And second module indicate the user review rating on user previously review data.

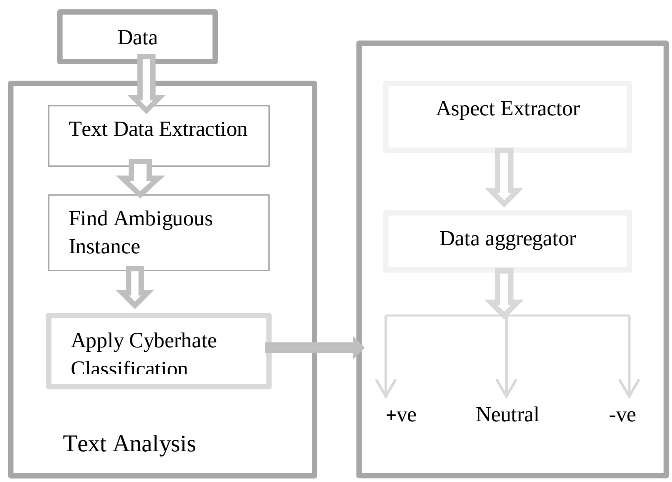

Fig 1: Proposed system architecture

This section describes the proposed fuzzy approach for cyberhate classification, in particular, introduce the theoretical preliminaries of fuzzy logic. A. Theoretical Preliminaries: It employs continuous truth values ranging from 0 to 1, rather than binary truth value (0 or 1). In the context of fuzzy sets, each of the elements e1,e2,...,en has a certain degree of membership to the set A, and the membership degree value depends on the membership function defined for the fuzzy set. In fuzzy rule-based systems, the main operation of fuzzification in the training stage can be done by transforming each numerical attribute into several qualitative attributes. B. Fuzzy Approach Methodology: Proposed fuzzy approach for cyberhate detection involves two steps the training stage, as illustrated in Fig.2.

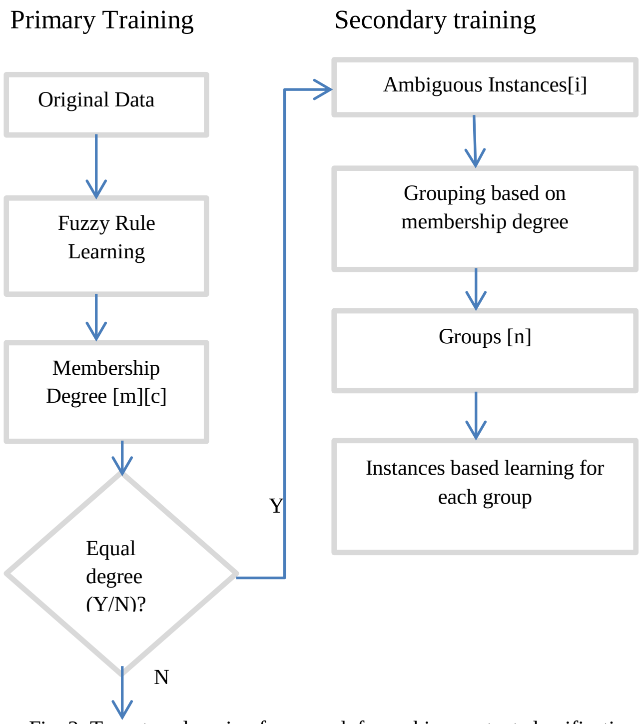

Fig. 2: Two stage learning framework for ambiguous text classification.

In this figure, i, c, and n represent the index of an instance, a class, and a classifier, respectively. In the first step, a set of fuzzy rules is trained using the mixed fuzzy rule formation algorithm. The procedure of the mixed fuzzy rule formation algorithm involves a sequential and constructive generation of new rules and modification of existing rules in an instance-by-instance manner. In the whole procedure, each rule involves membership functions. Once a new rule is added or an existing rule is modified after checking an instance, it is necessary to trigger the third case to avoid conflict of classification.

By using the above-mentioned procedure of the mixed fuzzy rule formation algorithm, a set of fuzzy rules is trained. In the testing stage, a new instance is classified through fuzzification, inference, and defuzzification. The fuzzification operation is simply aimed at mapping the numeric value of each feature of the new instance into a membership degree to each rule in each dimension. The defuzzification operation is to finally classify the new instance by assigning it the class with the maximum membership degree. Due to the text ambiguity, it is possible that an instance could obtain equal membership degrees for the hate and nonhate classes, which leads to increase in error when classifying unseen instances through a fixed rule. In this case here propose an approach to train the multiple fuzzy classifiers using different fuzzy norms to encourage diversity between these fuzzy classifiers. Therefore, the fusion of these fuzzy classifiers is likely to reduce the number of ambiguous instances in the training stage and has the chance to disambiguate an unseen instance that obtains equal membership degrees to both classes in the testing stage. The fusion can be achieved through averaging the overall membership degrees obtained from these fuzzy classifiers for each class. If so, the instance will be sent to the next stage for instance-based learning or instance-based reasoning, depending on whether it is a training instance or a test instance. However, the fusion of fuzzy classifiers cannot guarantee that all the ambiguous instances are disambiguated, so the second step is required to collate all the remaining ambiguous instances and produce a new training set. Overall, the proposed fuzzy approach involves two main stages. In the first stage, multiple fuzzy classifiers are trained using the mixed fuzzy rule formation algorithm alongside different fuzzy norms, and the fuzzy classifiers are then fused to identify ambiguous instances. In the second stage, the ambiguous instances are collated to produce the second training set for using KNN to classify new instances that are ambiguous [29].

IV. APPLICATION

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world.

Sentiment analysis can be used in diverse fields for various purposes. Few of them are follows

1. Online Commerce: The most general use of sentiment analysis is in ecommerce activities. Websites allows their users to submit their experience about shopping and product qualities. They provide summary for the product and different features of the product by assigning ratings or scores. Customers can easily view opinions and recommendation information on whole product as well as specific product features. Popular merchant websites like amazon.com provides review from editors and also from customers with rating information. http://tripadvisor.in is a popular website that provides reviews on hotels, travel destinations. They contain 75 million opinions and reviews worldwide. Sentiment analysis helps such websites by converting dissatisfied customers into promoters by analyzing this huge volume of opinions.

2. Voice of the Market (VOM): Voice of the Market is about determining what customers are feeling about products or services of competitors.

3. Voice of the Customer (VOC): Voice of the Customer is concern about what individual customer is saying about products or services.

4. Brand Reputation Management: Brand Reputation Management is concern about managing your reputation in market

5.Government policies: Sentiment analysis helps government in assessing their strength and weaknesses by analyzing opinions from public. Whether it is tracking citizens' opinions on a new rules, identifying strengths and weaknesses in a recruitment campaign in government job, assessing success of electronic submission of tax returns, or many other areas, we can see the potential for sentiment analysis. For example, the Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election. Being able to see the sentiment behind everything from forum posts to news articles means being able to strategies and plan for the future.

V. CONCLUSION

There is a huge need in the industry for such applications because every company wants to know how consumers feel about their products and services and those of their competitors. Sentiment analysis can be developed for new applications, the techniques and algorithms used for sentiment analysis have made good progress, but a lot of challenges in this field remain unsolved. More future research can be done for solving these challenges.

The work can be extended to build an advanced opinion mining system capable of rating the authenticity of a user review based on mining the sentiments or opinion threads of secondary reviewers. Several challenges still exist in the field of data classification and sentiment analysis and some of them are named entity recognition, co-reference resolution, domain dependency etc. These problems have to be tackled separately and those solutions can be used to improve sentiment analysis in near future.

REFERENCES

[1] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyber bullying on social networks based on bullying features," in the Proceedings of the 17th International Conference on Distributed Computer Network, pp. **4-7, Jan. 2016.**

[2] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in Proceedings of 10th International Conference on Machine Learning Application, Honolulu, USA, pp. **241-244,Dec. 2011.**

[3] J. H. Park and P. Fung, "*One-step and two-step classification for abusive language detection on Twitter,"* in Proc. 1st Workshop Abusive Lang. Online, Vancouver, BC, Canada, pp**. 41–45,Aug 2016.**

[4] H. Mubarak, K. Darwish, and W. Magdy, "*Abusive language detection on Arabic social media*," in Proceeding of 1st Workshop Abusive Lang. Online, Vancouver, BC, Canada, pp. **52–56,Aug 2017.**

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of Conference Empirical Methods Natural Lang. Process. (ACL), pp. **79-86,2002.**

[6] C. Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," in Proc. IEEE International Conference on Fuzzy Syst., Naples, Italy, pp**. 1–6,July 2017.**

[7] P. Burnap and M. L. Williams, *"Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making,"* Policy Internet, vol. **7**, issue **2**, pp. 223–242, **2015.**

[8] P. Burnap and M. L. Williams, *"Us and them: Identifying cyber hate on Twitter across multiple protected characteristics,*" EPJ Data Sci., vol. **5**, Issue **1**, pp. 11-15, **2016.**

[9] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, *"Sentiment analysis of Twitter data,*" in Proceedings of Workshop Language on Social Media, Stroudsburg, PA, USA, pp. 30–38,June **2011.**

[10] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Proceedings of 27th AAAI Conference on Artificial Intelligence, Bellevue, WA, USA, pp. **1621–1622,July 2013.**

[11] A. Mahmud, K. Z. Ahmed, and M. Khan, "Detecting flames and insults in text,*"* in Proeeding of 6th International Conference on Natural Language Processing., Gothenburg, Sweden, pp. **25–27, Aug 2008.**

[12] B. Gambäck and U. K. Sikdar, *"Using convolutional neural networks to classify hate-speech,"* in Proc. 1st Workshop Abusive Lang. Online, Vancouver, BC, Canada, Aug, pp. **85–90,Aug 2017.**

[13] D. Chandran, K. A. Crockett, D. Mclean, and A. Crispin, "An automatic corpus based method for a building multiple fuzzy word dataset," in Proceedings of IEEE International Conference on Fuzzy Syst., Istanbul, Turkey, pp. **1–8, Aug 2015.**

[14] I. B. Sassi, S. B. Yahia, and S. Mellouli, "Fuzzy classification-based emotional context recognition from online social networks messages," in Proc. IEEE Int. Conf. Fuzzy Syst., Naples, Italy, pp. **1–6,July 2017.**

[15] S. Asur and B. Huberman, "Predicting the future with social media*,"* in International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, vol. **1**, pp. 492–499,**2010.**

[16] D. Hardt and J. Wulff*, "What is the meaning of 5 *'s? An investigation of the expression and rating of sentiment,"* in Proceedings of KONVENS 2012, J. Jancsary, Ed. OGAI, pATHOS 2012 workshop,pp. 319-326, **2012.**

[17] H. Chen, P. De, Y. Hu, and B.-H. Hwang, *"Sentiment revealed in social media and its effect on the stock market,"* in Statistical Signal Processing Workshop (SSP), 2011 IEE, pp.**25-28, 2011.**

[18] S. P. Robertson, "Changes in referents and emotions over time in election-related social networking dialog," in System Sciences (HICSS), on 44th Hawaii International Conference, pp. **1–9, 2011.**

[19]M. Salath´e and S. Khandelwal*, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control,"* PLoS computational biology, vol. **7**, no. 10, **2011.**

[20] T. Menezes, C. Roth, J.-P. Cointet, "*Finding the semantic-level precursors on a blog network."* IJSCCPS, vol. **1**, pp. 115–134, **2011.**

[21] Han Liu, Pete Burnap, Wafa Alorainy, Matthew L.Williams, "A fuzzy approach to text classification with two stage training for ambiguous instances*"*, in proceeding of IEEE Transactions On Computational Social Systems ,vol. **6**,no.**2** ,pp. 227-240,**2019.**

[22] R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra,"*A review: Design and development of novel techniques for clustering and classification of data*", International journal of scientific research in computer sciences and engineering,Vol.**06**, Special Issue.**01** , pp.19-22, **Jan-2018.**

[23]V.K.Jin, N.Tripathi,"*Speech feature analysis and biometric person identification in multilingual environment"*, International journal of scientific research in network security and communication,Vol.**6**, Issue **1**,pp.7-11, **Feb 2011.**

[24]Bindushree V,Rashmi G.R, Uma H.R, *"Analysis of text recognition with data mining technique"*, International journal of scientific research in computer science and engineering,Vol.**7**,Issue.**6**,pp. 40-42,**Dec 2019.**

AUTHORS PROFILE

Miss.Supriya M. Yawalkar, she completed her B.E in Computer Science and Engineering from Rashtrasant Tukdoji Maharaj Nagpur University and currently pursuing M. E in Computer Science and Engineering from P.R.Pote (Patil) College of Engineering & Management from Sant Gadge Baba Amravati University ,Maharashtra, India. Her research interests include data mining.

Dr.A.S.Kapse, he completed his B.E in Computer Science and Engineering. He also achieved first class in M.E in Computer Science and Engineering. He Completed his Ph.D. and currently working as assistant professor at P.R.Pote (Patil) College of Engg & Management, Amravati, and Maharashtra, India. His subject specialization is Device Forensics. He has 11 years of teaching experience. He also worked in industry for the period of 1.5 year. He has 26 publications. He has published 2 papers in National and 7 papers in International conferences. He has professional membership in ISTE. He has interaction with professional institution such as DST, EDII.

