

A Survey on Hybrid Framework for Sentiment Analysis using Genetic Algorithm based Feature Reduction

Bhagyashree Bhoyar¹, Tejaswini Patil², Sonali Thosar³

^{1,2,3}Assistant Professor, Dept. of Computer Engg., Dr. D.Y. Patil Inst. of Engineering & Technology, Pune, India

Abstract: Sentiment analysis has become an important opinion mining technique due to the rapid development of Internet technologies and social media. Recent analysis work has represented the effectiveness of different sentiment classification techniques starting from easy rule-based and lexicon-based approaches to a lot of complicated machine learning algorithms. Whereas lexicon-based approaches have suffered from the lack of dictionaries and labeled knowledge, machine learning approaches have fallen short in terms of accuracy. To resolve the quantify ability issue that arises because the feature-set grows, a completely unique genetic rule (GA)-based feature reduction technique is projected. By using this hybrid approach, we have a tendency to be ready to scale back the feature-set size by up to forty-two while not compromising the accuracy. In future, the prediction of the applicability of our proposed work in various areas including security and surveillance, law-and-order, and public administration.

Keywords: Genetic rule, Sentiment analysis, classification technique.

1. Introduction

In today's era Internet and associated internet technologies have dramatically changed the means our society works. Social networks such as Facebook and Twitter became commonplace for exchanging concepts, sharing data, promoting business and trade, running political and philosophic campaigns, and promoting product and services. Social media is mostly studied from completely different views i.e., collection business intelligence for product and services promotion, watching malicious activities for sleuthing and mitigating cyber-threats, and sentiment analysis for analyzing people's feedback and reviews. Sentiment analysis, often referred as opinion mining, is that the extraction, identification, or characterization of the sentiment from text victimization Natural Language process (NLP), statistics, or machine learning (ML) ways. The field of sentiment analysis has been wide studied by researchers throughout the last few years. Sentiment analysis is generally mentioned within the context of product reviews like; is that this product review positive or negative? Are customers satisfied or dissatisfied? what is more, it additionally helps to answer the Business Intelligence connected queries like; Why aren't shoppers shopping for our product? but, cross-domain

insights and applications of sentiment analysis are scarce. The samples of such applications include analysis of user opinion on the politics, sociology, and the science of society. The proposes a hybrid approach to sentiment analysis which employs progressive milliliter algorithms and lexical databases to mechanically analyze archives of on-line documents (e.g., reviews, chats, and social media data). The propose a unique Genetic formula (GA) based mostly answer to feature reduction drawback by developing a bespoke fitness perform. The fitness perform utilizes SentiWord-Net lexicon to calculate the polarity distinction between a class label and possible feature vector (potential solution).

To the most effective of our information, we have a tendency to be the first to use such a hybrid approach with GA based mostly optimized feature selection. This organic process approach for optimum feature selection ends up in multiplied accuracy and higher quantifies ability.

The bespoke fitness perform shows up to forty-two reduced feature-set with none compromise on overall accuracy. Furthermore, so as to demonstrate the feasibility of the proposed feature reduction formula, to have a tendency to conjointly perform detailed comparison with different feature reduction algorithms including PCA and LSA which ends in system having up to fifteen.4% multiplied accuracy over PCA and up to 40.2% multiplied accuracy over LSA. PCA may be a spatiality reduction procedure that simplifies the complexity in high-dimensional knowledge by reducing an outsized set of variables to a small set that also retains data and trends gift in data. It comes a collection of points onto a smaller dimensional affine topological space of "best t". LSA may be a methodology employed in NLP that discovers a knowledge illustration that includes a lower dimension than the initial linguistics area by analyzing relationships between documents and its terms. It decreases the dimension employing a mathematical technique known as singular value decomposition (SVD).

2. Related Work

In this segment we talk about the conspicuous related research being completed in the zone of opinion investigation

and content mining. Our correlation criteria depends on the two variables to examined previously; combination of conclusion examination approaches in a bound together manner and a cross-disciplinary application zone. The intrigued to perceive how client's supposition and his/her social conduct can be useful in dissecting the current geopolitical circumstance and uprising.

Medhat et al. [18] presented a comprehensive summary of the recently planned algorithms, enhancements, and applications in the space of sentiment analysis. They additionally mentioned the connected fields to sentiment analysis e.g., transfer learning, feeling detection, and building resources. They tried to offer a full image of the sentiment analysis techniques and connected fields with transient details.

Khan et al. [19] proposed a rule-based domain-independent method which classifies subjective and objective sentences from reviews and blog comments. SentiWordNet is used to calculate the score and to determine the polarity. They showed that their proposed method is effective and it outperforms ML-based methods with an accuracy of 76.8% at the feedback level and 86.6% at the sentence level. Our proposed approach is aligned with these studies as we are also focusing on ML and lexicon-based methods. However, we are employing GA based optimized feature selection for training ML algorithms.

Agarwal et al. [20] inspected notion investigation on Twitter information. They presented POS-specific earlier extremity highlights furthermore, investigated the utilization of a tree bit to deter the requirement for dreary element building. Their new highlights furthermore, the tree piece performed nearly at the equivalent level and both beat the cutting edge standard strategies.

Kouloumpis et al. [21] examined the utility of etymological highlights for distinguishing the assessment of Twitter messages. They assessed the convenience of the current lexical assets just as the imaginative language utilized in microblogging.

[4] introduced a language-autonomous model for slant examination for short content structures e.g., informal organizations statuses. They utilized Twitter datasets to demonstrate glad and miserable opinions and demonstrated that their framework performed 10% superior to Naive Bayes (NB) model. These three papers are utilizing feeling investigation on short-content information i.e., SMS, tweets and so forth. So also, Pontiki et al. [22] portrayed the angle based supposition examination. They identified the parts of given target elements and the assumption communicated for every angle. They utilized physically commented on surveys of eateries and workstations as a dataset.

Njolstad et al. [23] proposed, defined, what's more, assessed four distinctive element classifications formed of 26 article highlights for notion examination. They utilized diverse ML techniques to prepare conclusion classifier of Norweign financial web news articles. They accomplished classification

exactness up to 71%. When contrasting ML classifiers, they found that J48 yielded the most elevated execution intently pursued by Random Forest (RF). TO have likewise introduced a comparative examination in which analyzed diverse classifiers and their precision on framework. In any case, expanded our assessment by including GA enhanced includes in correlation.

Govindarajan [24] proposed a half and half classification strategy in light of incorporation classification techniques utilizing arcing classifier. They broke down the execution as far as exactness. They planned classifier group utilizing NB and GA. They assessed the adequacy of group procedure for notion examination. At long last, they assessed the execution under distinctive execution measurements utilizing film audits datasets. Be that as it may, they don't analyze the execution of various classifiers and don't give any enhancement to include estimate decrease.

As see that the greater part of the related work utilized autonomous systems for assessment investigation while utilizing barely any assessment measurements. Besides, they don't give the client with the opportunity to pick diverse calculations, classifiers, and improvements as per redid needs. Interestingly, our proposed structure crosses over any barrier between feeling examination and geopolitical knowledge by giving

- A unified system having the office to plug unique calculations, cross-approval, and improved element choice
- A two-dimensional examination on popular sentiments in affiliation with political uprisings by consolidating security and conclusion mining.

3. Conclusion

We Concluded that our opinion investigation system has turned out to be a brilliant expansion inside the control of feeling mining. It gave the edibility of choosing among 3 generally utilized opinion investigation systems in accordance with custom needs. With further benefits of GA principally based enhancement, it diminishes include measure and improves efficiency while keeping up the quantify ability. inside the future, we tend to plan to build this structure for digital insight all together that it may encourage create suggestions for law-authorization offices upheld client conclusions.

References

- [1] P. DiMaggio, E. Hargittai, W. R. Neuman, and J. P. Robinson, "Social implications of the Internet," *Annu. Rev. Sociol.*, vol. 27, pp. 307-336, Aug. 2001.
- [2] C. Wang and P. Zhang, "The evolution of social commerce: The people, management, technology, and information dimensions," *Commun. Assoc. Inf. Syst.*, vol. 31, no. 5, pp. 1-23, 2012.
- [3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval* 2(1-2), pp. 1-135, 2008.

- [4] A. Davies and Z. Ghahramani, "Language-independent Bayesian sentiment mining of Twitter," in Proc. Workshop Social Netw. Mining Anal., 2011, pp. 99-107.
- [5] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *J. Informetrics*, vol. 3, no. 2, pp. 143-157, 2009.
- [6] A. Collomb, C. Costea, D. Joyeux, O. Hasan, and L. Brunie, "A study and comparison of sentiment analysis methods for reputation evaluation," Tech. Rep. RR-LIRIS-2014-002, 2014.
- [7] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526-558, 2009.
- [8] M. Taboada, J. Brooke, M. Toloski, K. Voll, and M. Stede, "Lexicon based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267-307, 2011.
- [9] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 231-240.
- [10] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82-89, 2013.
- [11] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in Proc. LREC, vol. 6, 2006, pp. 417-422.
- [12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1-3, pp. 37-52, 1987.
- [13] S. T. Dumais, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.* vol. 38, no. 1, pp. 188-230, 2004.
- [14] S. Goel, "Cyberwarfare: Connecting the dots in cyber intelligence," *Commun. ACM*, vol. 54, no. 8, pp. 132-140, Aug. 2011.
- [15] UCI ML Repository Sentiment Analysis Dataset, 2015. <http://archive.ics.uci.edu/ml/datasets/Sentiment+Labeled+Sentences>
- [16] J. A. Bowden. (2016). Twitter Sentiment Analysis. <https://old.datahub.io/dataset/twitter-sentimentanalysis>
- [17] J. Littman, L. Wrubel, and D. Kerchner, 2016 United States Presidential Election Tweet IDS, 2016.
- [18] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093-1113, 2014.
- [19] A. Khan, B. Baharudin, and K. Khairullah, "Sentiment classification using sentence-level lexical based semantic orientation of online reviews," *Trends Appl. Sci. Res.*, vol. 6, no. 10, pp. 1141-1157, 2011.
- [20] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in Proc. Workshop Lang. Social Media, 2011, pp. 30-38.
- [21] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the OMG!," in Proc. ICWSM, vol. 11. 2011, pp. 538-541.
- [22] M. Pontiki et. al., "SemEval-2016 task 5: Aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), 2014, pp. 27-35.
- [23] P. C. S. Njølstad, L. S. Høysæter, W. Wei, and J. A. Gulla, "Evaluating feature sets and classifiers for sentiment analysis of financial news," in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT), vol. 2, Aug. 2014, pp. 71-78.
- [24] M. Govindarajan, "Sentiment analysis of movie reviews using hybrid method of naive Bayes and genetic algorithm," *Int. J. Adv. Comput. Res.*, vol. 3, no. 4, pp. 139-145, 2013.
- [25] A. McCallum. (1998). Rainbow Stopwords. <http://www.cs.cmu.edu/~mccallum/bow/rainbow/>
- [26] M. McCandless, E. Hatcher, and O. Gospodnetic, "Lucene in Action," Second Edition, July 2010.