Text Generation Models and Imbalanced Sentiment Analysis

Subjects: Computer Science, Artificial Intelligence

Contributor: Cici Suhaeni , Hwan-Seung Yong

The significance of sentiment analysis has extended across a wide range of fields, finding extensive use in various applications. As digital communication continues to expand, the ability of sentiment analysis to interpret complex human emotions and opinions becomes increasingly important, proving invaluable in fields ranging from social sciences to customer service and beyond. In this era of increasing digitization, leveraging the power of data through sentiment analysis offers unique insights, making significant contributions to sectors such as those previously summarized in various studies, namely, healthcare, social policy, e-commerce, and digital humanities.

sentiment classification synthetics review generation text classification text generation synthetic data generation natural language processing sentiment analysis imbaqlanced sentiment analysis

1. Text Generation

A recent systematic literature review by Fatima et al. [1] scrutinized 90 primary studies conducted from 2015 to 2021, which highlighted methods for generating text, quality measures, datasets, and languages, along with their usage in the context of deep learning. This research emphasized the escalating interest in deep learning methodologies for text generation over the studied period. Significantly, it highlighted the potential of GPT-3 in generating text due to its extensive training and substantial generative capabilities. Iqbal and Qureshi [2] furthered this by demonstrating that current deep learning methods applied in the realm of synthetic text creation encompass Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs).

GAN-based text generation has been explored extensively in recent studies. Wang and Wan [3] unveiled a fresh architectural framework—SentiGAN, which encompasses multiple generators and one multi-class discriminator, all architected specifically to concoct a wide range of examples that all carry a specific sentiment label. Building on this, Liu et al. [4] advanced this framework by proposing a GAN that is aware of its category (CatGAN). This was equipped with an efficient model for generating text according to its category, in addition to a hierarchical algorithm for evolutionary learning dedicated to training the model.

The revolutionary “Transformer” model was introduced by Vaswani et al. [5], providing the groundwork for subsequent language generation models, including the GPT and BERT architectures. Following this, Radford et al.

[6] presented a seminal paper introducing the GPT-2 model, a noteworthy development in the field of language generation.

Recent studies have employed GPT-2 in various innovative ways for text generation. Anaby-Tavor et al. [7] leveraged GPT-2 in a method called LAMBADA, while Ma et al. [8] proposed the Switch-GPT method. Xu et al. [9] used GPT-2 and T5 to generate table captions, and Bayer et al. [10] also utilized GPT-2, but because of some limitations, they suggested GPT-3 as a viable choice for enhancing results, having utilized GPT-2 in their proposed method for text generation.

The introduction of GPT-3, the successor of GPT-2, marked another milestone in this field [11]. Recently, Zhong et al. [12] investigated the understanding ability of ChatGPT, a GPT model variant, by subjecting it to the well-known GLUE benchmark test and juxtaposing its performance against four emblematic models that had been fine-tuned in the style of BERT. These studies form the backbone of researchers' understanding and application of text generation and sentiment analysis, with this research intending to contribute further to this growing body of knowledge.

2. Imbalanced Sentiment Analysis

The topic of imbalanced sentiment analysis has been a vibrant area of research in recent years, with numerous approaches developed to tackle this problem.

Obiedat et al. [13] introduced a combined method that melds the Support Vector Machine (SVM) algorithm with Particle Swarm Optimization (PSO), along with several oversampling methods to tackle the problem of unbalanced sentiment analysis within a dataset of customer reviews. This tactic proved successful in dealing with data disparity, showcasing the promise of these hybrid methods in this field.

Han Wen and Junfang Zhao [14] introduced an alternate strategy, which suggested a technique for sentiment evaluation of unbalanced comment data utilizing a BiLSTM structure. The approach involved Adaptive Synthetic Sampling in cases where the dataset contained more negative instances than positive ones, deploying a model based on CNN-BiLSTM for classifying the sentiment.

In the same spirit, Tan et al. [15] crafted an innovative hybrid system that amalgamates the advantages of the Transformer model, exemplified by the Robustly Optimized BERT Pretraining Approach (RoBERTa), and the Recurrent Neural Network, embodied by Gated Recurrent Units (GRUs). This hybrid system was engineered to address the issue of unbalanced datasets by applying data augmentation via word embeddings, while oversampling the minority class, thereby boosting the model’s ability to represent data and its resilience in executing sentiment classification tasks.

Wu and Huang [16] proposed a different method for handling imbalanced text data. They introduced a hybrid method, which utilizes a generative adversarial network alongside the Shapley algorithm, termed HEGS. This

structure could produce a wide range of training phrases to level the textual data and bolster the ability to classify instances belonging to the minority classes.

Almuayqil et al. [17] took an innovative approach by designing a model specifically for imbalanced Twitter datasets. By utilizing an array of text sequencing preprocessing methods combined with random under-sampling of the majority class, they managed to considerably cut down the computational time required for the task.

Further investigating Twitter data, Ghosh et al. [18] assessed the efficacy of varying proportions of synthetic oversampling techniques to manage class imbalance in Twitter sentiment analysis. Concurrently, George [19] introduced a unique synthetic oversampling method, SMOTE, amalgamated with a composite model referred to as the Ensemble Bagging Support Vector Machine (EBSVM), to address the problem of data imbalance.

Cai and Zhang [20] adopted a unique perspective by concentrating on sentiment information extraction from an imbalanced short text review dataset. They introduced a fusion multi-channel BLTCN-BLSTM self-attention sentiment classification strategy, amalgamating focus loss rebalancing and classifier enhancement mechanisms to boost sentiment prediction accuracy.

A recent approach to handling imbalanced sentiment analysis is by generating artificial text for minority classes. Imran et al. [21] utilized a GAN-based model to generate synthetic data for tackling this problem. Similarly, Habbat et al. [22] employed a pretrained AraGPT-2-based model to create synthetic Arabic text, addressing the issue of imbalanced sentiment analysis. Following this, they utilized AraBERT for textual representation and a deep learning model stack for classification. This research illuminates the potential of language-specific models in proficiently managing tasks related to imbalanced sentiment analysis.

Lastly, Ekinci [23] performed a comparative study of imbalanced offensive data classification using an LSTM-based sentence generation method. Various classifiers were trained using TF-IDF and Word2vec for text representation, demonstrating the value of sentence generation methods in handling imbalanced sentiment analysis tasks.

Together, these studies highlight the diverse methods and models available to handle imbalanced sentiment analysis, and they set the foundation for further research in this field.

References

1. Fatima, N.; Imran, A.S.; Kastrati, Z.; Daudpota, S.M.; Soomro, A. A Systematic Literature Review on Text Generation Using Deep Neural Network Models. IEEE Access 2022, 10, 53490–53503.
2. Iqbal, T.; Qureshi, S. The Survey: Text Generation Models in Deep Learning. J. King Saud. Univ. Comput. Inf. Sci. 2022, 34, 2515–2528.

3. Wang, K.; Wan, X. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm Sweden, 13–19 July 2018; pp. 4446–4452.
4. Liu, Z.; Wang, J.; Liang, Z. CatGAN: Category-Aware Generative Adversarial Networks with Hierarchical Evolutionary Learning for Category Text Generation. Proc. AAAI Conf. Artif. Intell. 2020, 34, 8425–8432.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need 2023. Available online: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 31 July 2023).
6. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. OpenAI Blog 2019, 1, 9.
7. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Not Enough Data? Deep Learning to the Rescue! arXiv 2019.
8. Ma, C.; Zhang, S.; Shen, G.; Deng, Z. Switch-GPT: An Effective Method for Constrained Text Generation under Few-Shot Settings (Student Abstract). Proc. AAAI Conf. Artif. Intell. 2022, 36, 13011–13012.
9. Xu, J.H.; Shinden, K.; Kato, M.P. Table Caption Generation in Scholarly Documents Leveraging Pre-Trained Language Models. In Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 12–15 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 963–966.
10. Bayer, M.; Kaufhold, M.-A.; Buchhold, B.; Keller, M.; Dallmeyer, J.; Reuter, C. Data Augmentation in Natural Language Processing: A Novel Text Generation Approach for Long and Short Text Classifiers. Int. J. Mach. Learn. Cybern. 2023, 14, 135–150.
11. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. arXiv 2020.
12. Zhong, Q.; Ding, L.; Liu, J.; Du, B.; Tao, D. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-Tuned BERT. arXiv 2023.
13. Obiedat, R.; Qaddoura, R.; Al-Zoubi, A.M.; Al-Qaisi, L.; Harfoushi, O.; Alrefai, M.; Faris, H. Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution. IEEE Access 2022, 10, 22260–22273.
14. Wen, H.; Zhao, J. Sentiment Analysis Model of Imbalanced Comment Texts Based on BiLSTM. In Review: 2023. Available online: https://www.researchsquare.com/article/rs-2434519/v1 (accessed on 31 July 2023).

15. Tan, K.L.; Lee, C.P.; Lim, K.M. RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. Appl. Sci. 2023, 13, 3915.
16. Wu, J.-L.; Huang, S. Application of Generative Adversarial Networks and Shapley Algorithm Based on Easy Data Augmentation for Imbalanced Text Data. Appl. Sci. 2022, 12, 10964.
17. Almuayqil, S.N.; Humayun, M.; Jhanjhi, N.Z.; Almufareh, M.F.; Khan, N.A. Enhancing Sentiment Analysis via Random Majority Under-Sampling with Reduced Time Complexity for Classifying Tweet Reviews. Electronics 2022, 11, 3624.
18. Ghosh, K.; Banerjee, A.; Chatterjee, S.; Sen, S. Imbalanced Twitter Sentiment Analysis Using Minority Oversampling. In Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 23–25 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
19. Department of Computer Science; Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India; George, S.; Srividhya, V. Performance Evaluation of Sentiment Analysis on Balanced and Imbalanced Dataset Using Ensemble Approach. Indian J. Sci. Technol. 2022, 15, 790–797.
20. Cai, T.; Zhang, X. Imbalanced Text Sentiment Classification Based on Multi-Channel BLTCN-BLSTM Self-Attention. Sensors 2023, 23, 2257.
21. Imran, A.S.; Yang, R.; Kastrati, Z.; Daudpota, S.M.; Shaikh, S. The Impact of Synthetic Text Generation for Sentiment Analysis Using GAN Based Models. Egypt. Inform. J. 2022, 23, 547–557.
22. Habbat, N.; Nouri, H.; Anoun, H.; Hassouni, L. Using AraGPT and Ensemble Deep Learning Model for Sentiment Analysis on Arabic Imbalanced Dataset. ITM Web Conf. 2023, 52, 02008.
23. Ekinci, E. Classification of Imbalanced Offensive Dataset—Sentence Generation for Minority Class with LSTM. Sak. Univ. J. Comput. Inf. Sci. 2022, 5, 121–133.