Azerbaijan Journal of High Performance Computing, Vol 6, Issue 2, 2023, pp. 191-198
https://doi.org/10.32010/26166127.2023.6.2.191.198

Unlocking Educational Insights: Integrating Word2Vec Embeddings and Naive Bayes Classifier for Serious Game Data Analysis and Enhancement

Anar Mammadli

Azerbaijan State Oil and Industry University, Baku, Azerbaijan, anar.mammadli.az@asoiu.edu.az

*Correspondence: Anar Mammadli, Azerbaijan State Oil and Industry University, Baku, Azerbaijan, anar.mammadli.az@asoiu.edu.az

Abstract

This study explores the integration of Word2Vec embeddings and machine learning models to analyze and enhance serious game data. Word2Vec captures semantic relationships in textual content, while the Naive Bayes classifier extracts meaningful patterns. The approach improves understanding of linguistic nuances, contributing to the effectiveness of serious3 games in achieving educational objectives. Experimental results demonstrate the model's efficacy in uncovering hidden insights within the game data. This research provides a robust framework for optimizing serious game content and enhancing its educational impact.

Keyword: Serious Game, Artificial Intelligence, NLP, Text Categorization, Embeddings

1. Introduction

Serious games, designed for educational and training purposes, have emerged as powerful tools to engage and motivate learners. These games leverage interactive and immersive experiences to facilitate learning in diverse domains. Understanding serious game data's underlying patterns and semantic structures is crucial for optimizing educational outcomes. This study investigates the integration of Word2Vec embeddings and Naive Bayes classifier models to analyze and enhance serious game data. This approach aims to unravel the intricate linguistic nuances embedded in serious games by transforming textual content into vector representations and extracting meaningful patterns. The exploration of such methodologies is pivotal for advancing the design and effectiveness of serious games in educational technology.

Word2Vec embeddings play a significant role in analyzing textual data within the context of Serious Games designed for educational and training purposes. These embeddings are instrumental in transforming raw textual content into numerical representations, thereby enabling extracting meaningful patterns and semantic structures. Integrating Word2Vec embeddings is a crucial aspect of optimizing educational outcomes and enhancing the overall effectiveness of serious games in educational technology.

This scientific paper explores the integration of Word2Vec embeddings and Naive

Bayes classifier as a robust methodology for analyzing and enhancing serious game data. The transformative power of Word2Vec lies in its ability to convert textual content into vector representations, unveiling intricate semantic structures within the linguistic fabric of serious games. When coupled with the Naive Bayes classifier, this fusion promises to unravel nuanced patterns, fostering a deeper understanding of how learners interact with educational content. The insights garnered through this approach can potentially revolutionize the design and effectiveness of serious games within the ever-evolving landscape of educational technology. This paper delves into the methodologies employed, the significance of Word2Vec embeddings, and the implications of using the Naive Bayes classifier, aiming to contribute novel perspectives to the intersection of linguistics, technology, and educational science.

2. Literature Review

The work introduced by (Marwa et al., 2017) reveals that, regardless of the language used, negative sampling emerges as the most efficient algorithm for Word2Vec in the context of topic segmentation. However, the choice of learning models requires careful consideration, as Continuous Bag of Words (CBOW) demonstrates higher efficiency with frequent words, while Skip-Gram excels with infrequent words. Compared to LSA and GloVe, Word2Vec and GloVe exhibit superior effectiveness, with Word2Vec showcasing the best word vector representations, particularly in a small-dimensional semantic space. In a comprehensive comparison, we establish that Word2Vec and GloVe outperform LSA in terms of effectiveness for topic segmentation. Furthermore, work demonstrates that Word2Vec excels over GloVe, particularly when considering the dimensionality of the semantic space.

(Pennington et al., 2014) addresses the ongoing debate between prediction-based and count-based models for word representation learning. While prediction-based models, exemplified by Baroni et al. (2014), have gained significant support, the authors argue that both classes of methods share fundamental similarities as they probe the underlying co-occurrence statistics of a corpus. The key distinction lies in the efficiency with which count-based methods capture global statistics. The authors propose a novel model, GloVe, which combines the advantages of count data efficiency with the ability to capture linear substructures found in recent prediction-based methods like word2vec. GloVe, a global log-bilinear regression model, is designed for unsupervised learning of word representations and demonstrates superior performance on various tasks, including word analogy, word similarity, and named entity recognition, compared to other existing models.

In this paper (Altszyler et al., 2017), the authors compare the efficacy of Skip-gram and Latent Semantic Analysis (LSA) in learning word embeddings from small text corpora. Their evaluation involves testing the models' ability to represent semantic categories in nested subsamples of a medium-sized corpus. The results indicate that Word2Vec embeddings outperform LSA when trained with medium-sized datasets (approximately 10 million words).

However, in scenarios with reduced corpus sizes, Word2Vec's performance significantly declines, making LSA a more suitable tool.

Several works related to machine learning have been done in the serious games area. Such as feature selection methods (Azam & Yao., 2011) for serious game data, classification of textual data [5], and automated NPC behavior generation (Dobrovsky et al., 2017; Serafim et al., 2017; Jeerige et al., 2019) from serious games. Classification also applied to textual datasets in these papers (Azam & Yao., 2011; Mammadli & Ismayilov., 2023; Zagal et al., 2005) using Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN). Extensive investigations have been undertaken to enhance the efficacy of classification techniques across diverse applications, including their integration into serious games. Despite the acknowledged significance of classification in serious gaming scenarios, the exploration of Graph Neural Networks (GNNs) in this context remains comparatively limited.

3. Methodology

3.1. Preprocessing

The research endeavors to explore and analyze a specialized dataset comprising words intricately linked to a central theme. In this dataset, each entry features a main word surrounded by others that share various semantic relationships, such as synonyms, antonyms, similar words, and those evoking the essence of the main 3word. Furthermore, the dataset includes supplementary information, which will be excluded for the purposes of the experiment. The methodology adopted for this investigation involves employing word2vec as an initial step, followed by applying the Naive Bayes classifier. This multifaceted approach aims to unravel intricate word associations and leverage the power of neural networks to extract meaningful patterns from textual data. Through these techniques, the study seeks to gain valuable insights into the interconnected nature of words within the dataset.

The research commences with the acquisition of raw data in CSV format, which is subsequently subjected to preprocessing involving the removal of superfluous columns. The retained columns encompass the main word, five associated words, and the corresponding category. Following the extraction of this pertinent information, textual data transformation ensues, and the CBOW (Zhang et al., 2010) and Skip-Gram models are applied. It is noteworthy to highlight the distinctions between CBOW and Skip-Gram, two prominent algorithms in word embedding techniques. The choice between the two models is made judiciously, selecting the one that aligns optimally with the intrinsic characteristics of the dataset under investigation. This methodological approach is designed to enhance the understanding of intricate word associations within the context of the dataset.

Raw data contains the main word and five different words that are related to this main word. Also, we have a category that explains the general category of the main word. These categories include item, food, human, location, animal, profession, or other. Raw data is shown in Figure 1.

	word_tabu	first_word	second_word	...	fourth_word	fifth_word	category
0	QƏLƏM	KAĞIZ	QƏLƏMQABI	...	ÇANTA	KİTAB	item
1	FİL DİŞİ	HEYVAN	BOZ	...	32	ÜZV	other
2	MAŞIN	SÜRMƏK	TƏKƏR	...	YOL	NƏQLİYYAT	item
3	TELEFON	EKRAN	İNTERNET	...	TUŞ	SMS	item
4	HƏKİM	MÜAYİNƏ	ƏMƏLİYYAT	...	XƏSTƏ	PEŞƏ	profession
..	...	...	...	...	...	...	...

Fig. 1: The first five rows from the data source

This raw data should be converted to a dataset to apply word2vec and our model. Firstly we convert this data to a pandas frame to do quick operations. After that, we explored categories and analyzed them. In Fig 2, there are categories that we want to keep shown. Still, we have an imbalanced dataset.

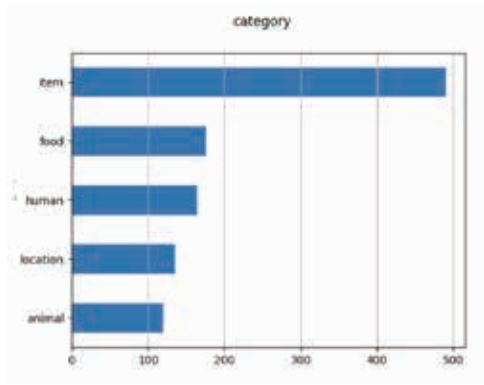

Fig. 2: Categories per number of data

Words similar to each other would be placed closer together to each other. It helps to understand how words are distributed according to each other. It helps to understand the similarity of words. You can see this in Figure 3.

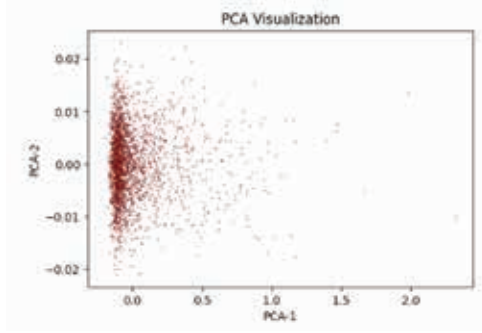

Fig. 3: PCA visualization for similarity of words

3.2. Word2Vec - Bag of Words

The bag-of-words (BOW) model is a method of converting any text into fixed-length vectors. It achieves this by tallying the frequency of each word present in the text, a procedure commonly known as vectorization. The structure closely resembles that of a feed-forward neural network. This model architecture seeks to anticipate a target word based on a set of context words. The underlying idea is straightforward: for a phrase like "Gözəl bir gün keçirin" we designate "gün" as the target word, with "gözəl," "bir," and "keçirin" as the context words. The model utilizes the distributed representations of these context words to predict the target word effectively.

3.3. Model

We employ the Naive Bayes algorithm for the model training, a sophisticated probabilistic classifier leveraging Bayes' Theorem (Vikramkumar et al., 2014). This theorem, rooted in probability theory, facilitates predictions by drawing upon prior knowledge of potentially correlated conditions. The Naive Bayes algorithm proves exceptionally apt for our dataset, evaluating each feature in isolation. It meticulously computes the probability associated with each category, culminating in the prediction of the category boasting the highest probability. This methodology, grounded in independence and precision, underscores the algorithm's suitability for our dataset's nuanced characteristics.

The classifier relies on Bayes' theorem, which is expressed as:

$$(C|X) = \frac{P(X|Ck)\cdot P(Ck)}{P(X)}$$

Where:

P(Ck|X) is the posterior probability of class Ck given the features X,

P(X|Ck) is the likelihood of observing the features X given class Ck,

P(Ck) is the prior probability of class Ck,

P(X) is the probability of observing the features X.

Now, applying the "naive" assumption of feature independence, the likelihood term can be expressed as the product of the individual feature probabilities:

$$P(X|Ck) = P(x1|Ck) \cdot P(x2|Ck) \cdot \ldots \cdot P(xn|Ck)$$

where x1, x2, …, xn are the individual features.

Assuming a document classification scenario with features representing terms or words, we can rewrite this as:

$$P(X|Ck) = P(w1|Ck) \cdot P(w2|Ck) \cdot \ldots \cdot P(wn|Ck)$$

Where w1, w2, …, wn are the terms in the document.

The classifier assigns a document to the class that maximizes the posterior probability, which can be expressed as:

$$\hat{y} = \text{argmax}_k P(Ck|X)$$

In practice, it is expected to work with logarithmic probabilities due to computational convenience and avoiding numerical underflow issues. Therefore, the decision rule becomes:

$$\hat{y} = \text{argmax}_k \log(P(Ck|X))$$

This involves computing the log-likelihoods of each term for each class and adding

them up with the logarithm of the prior probability for each class.

In summary, the Naive Bayes classifier employs Bayes' theorem, assuming feature independence, to calculate the probability of a document belonging to a particular class. The class with the highest probability is then assigned as the predicted class.

4. Results

The evaluation involves assessing its performance using various metrics. These metrics include accuracy, which measures the proportion of correct predictions; a confusion matrix providing a breakdown of correct and incorrect predictions by class; and recall, measuring the fraction of relevant instances that were successfully retrieved from the total amount. Accuracy, Precision, Recall, F1 score, and Support were used as evaluation criteria. F1-score was taken as the primary evaluation criterion since the training set is very imbalanced.

Experiments

For experiments, Serious Game data was used. This dataset consists of 1000 rows and 5 columns. The dataset is split into 70% train and 30% test.

Table 1.

	precision	recall	f1-score	support
Animal	1.00	0.55	0.71	33
Food	0.84	0.70	0.76	53
Human	1.00	0.16	0.27	51
Item	0.58	0.98	0.72	148
Location	1.00	0.10	0.18	41
Macro average	0.88	0.50	0.53	326
Weighted average	0.78	0.65	0.59	326
Accuracy			0.65	326

The Bag of Words (BoW) model got 65% of the test set right but struggles to recognize categories other than Item. The dataset we used is a small part of serious game data we used which is why it is hard to predict some categories. The confusion matrix in Figure 4. depicts that 145 out of 148 item categories are found correctly, but in other categories, it is not that precise. In our dataset, many items are related to humans, so it is hard to predict. It means that in the human category, there can be the main word human, which has an item that explains that humans and our model cannot understand that.

Conclusion and Future Works

In the results section, it is imperative to emphasize the necessity of further exploration and refinement of our dataset. A scientific approach dictates that we meticulously

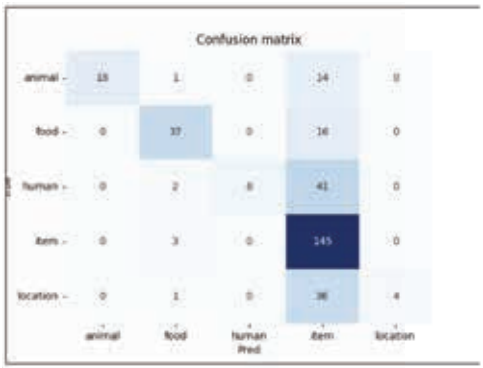

Fig. 4: Confusion Matrix

scrutinize the dataset, identifying potential areas for improvement and optimization. To enhance the robustness of our findings, we propose systematically experimenting with various models. By rigorously testing and comparing different models, we can gain insights into their strengths and weaknesses. This methodical exploration ensures that our conclusions are grounded in a comprehensive understanding of the dataset, ultimately contributing to the credibility and reliability of our study. The results underscore the potential for refinement and enhancement in our experimental approach. This suggests an opportunity for further investigation and improvement in the ongoing experiment. By identifying areas of potential optimization, we can iteratively fine-tune our methodology to yield more robust and reliable outcomes. This acknowledgment of room for improvement serves as an invitation to delve deeper into the experiment, fine-tune variables, and explore alternative avenues that may lead to heightened accuracy and efficiency in our model. This commitment to continuous improvement aligns with the dynamic nature of scientific inquiry and paves the way for a more comprehensive and impactful study.

References

Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520.*

Azam, N., & Yao, J. (2011, June). Incorporating game theory in feature selection for text categorization. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing* (pp. 215-222). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bayes, T. (1968). Naive bayes classifier. *Article Sources and Contributors,* 1-9.

Dobrovsky, A., Borghoff, U. M., & Hofmann, M. (2017). Applying and augmenting deep reinforcement learning in serious games through interaction. *Periodica Polytechnica Electrical Engineering and Computer Science, 61*(2), 198-208.

Georgios N.. Yannakakis, & Togelius, J. (2018). *Artificial Intelligence and Games.*

Springer.

Jeerige, A., Bein, D., & Verma, A. (2019, January). Comparison of deep reinforcement learning approaches for intelligent game playing. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0366-0371). IEEE.

Mammadli, A., & Ismayilov, E. A. (2023, August). Application of Deep Learning Technologies in Serious Games. In *2023 5th International Conference on Problems of Cybernetics and Informatics (PCI)* (pp. 1-4). IEEE.

Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science, 112,* 340-349.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Serafim, P. B. S., Nogueira, Y. L. B., Vidal, C., & Cavalcante-Neto, J. (2017, November). On the development of an autonomous agent for a 3d first-person shooter game using deep reinforcement learning. In *2017 16th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (pp. 155-163). IEEE.

Zagal, J. P., Mateas, M., Fernández-Vara, C., Hochhalter, B., & Lichti, N. (2005, June). Towards an ontological language for game analysis. In *DiGRA Conference* (pp. 1-13).

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics, 1,* 43-52.

Submitted 28.09.2023
Accepted 21.11.2023