# Deep Learning Framework for Identifying Future Market Opportunities from Textual User Reviews


Jordan Kralev

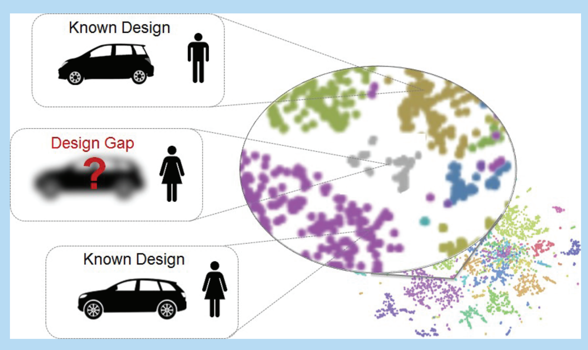
Department of Systems and Control / Technical University of Sofia
Digital Computational Linguistics / Bulgarian Academy of Sciences
jkralev@ieee.org



## ABSTRACT

The paper develops an application of design gap theory for identification of future market segment growth and capitalization from a set of customer reviews for bought products from the market in a given past period. To build a consumer feature space, an encoded-decoder network with attention is trained over the textual reviews after they are pre-processed through tokenization and embedding layers. The encodings for product reviews are used to train a variational auto encoder network for representation of a product feature space. The sampling capabilities of this network are extended with a function to look for innovative designs with high consumer preferences, characterizing future opportunities in a given market segment. The framework is demonstrated for processing of Amazon reviews in consumer electronics segment.


The design gap models allow the prediction of consumer preference for an "unknown and not existing products". The prediction cannot exactly tell what will be these future products. The model output is a bounded subset of the design space or feature space, which will be favored by the customers. Such bounded subset can be contrasted with the unbounded set of all possible future designs.




## MATHEMATICAL FORMULATION

Language Transformer

Encoder

$r = (r_1, r_2, r_3, \dots, r_L)^T$ Input text with tokens r and length L

$r_i^{\text{emb}} = E^T \mathbf{1}(r_i),$ Representation of tokens with embedding vectors

$$x_{ff,i+1} = F_{LSTM,ff}\left(x_{ff,i}, \begin{pmatrix} r_i^{\text{emb}} \\ e_{fb,i} \end{pmatrix}\right)$$
$$e_i = G_{LSTM,ff}\left(x_{ff,i}, \begin{pmatrix} r_i^{\text{emb}} \\ e_{fb,i} \end{pmatrix}\right)$$

Forward LSTM layer with two inputs
- token embedding
- feedback vector

$$x_{fb,i+1} = F_{LSTM,fb}(x_{fb,i}, e_i)$$
$$e_{fb,i} = G_{LSTM,fb}(x_{fb,i}, e_i)$$

Feedback LSTM layer

Decoder

$y = (y_1, y_2, y_3, \dots, y_L)^T \; y_i^{\text{emb}} = E^T \mathbf{1}(y_i),$ Recovered input sequence from decoder

$$x_{i+1} = F_{LSTM,D}\left(x_i, \begin{pmatrix} y_i^{\text{emb}} \\ c_i \end{pmatrix}\right)$$
$$d_i = G_{LSTM,D}\left(x_i, \begin{pmatrix} y_i^{\text{emb}} \\ c_i \end{pmatrix}\right)$$

Forward LSTM layer with two inputs
- token embedding
- context vector

$c_i = \sum_{j=1}^{L} a_{i,j} e_j.$ Context is weighted sum over encoded sequence

$a_i = \sigma(s_i) \quad a_i = (a_{i1,}, a_{i,2}, \dots, a_{i,L}),$ Weights vector is normalized to unit length

$s_{i,j} = V(W_1 e_j + W_2 x_i),$ Attention is function on input sequence biased with current decoder state

$\mu_i = \sigma(E W_3(d_i))$ Probability over embedding vocabolary

$y_{i+1} = \underset{z}{\operatorname{argmax}}\, \mu_i(z).$ Maximum likelihood selection for decoder token

Training cost function

$J_{LT} = -\sum_{i=1}^{L} E_{r_i}(ln(\mu_i)),$ Cross entropy between input and decoded sequence

Market Transformer

$\hat{X}_P = (e_L(r^1), e_L(r^2), \dots e_L(r^{N(P)})),$ Product category matrix from final encoder states

$\hat{X}_C = (X_{u,1}, X_{u,2}, \dots, X_{u,N(C)}),$ Client category matrix from final encoder states

Product feature space mean

Product feature space variance

$s_p \quad N(\mu, \sigma)$ Sampling normal distribution with given threshold

$\hat{y}_P = W_{dec}(s_p)$ Mapping from feature space to input space

Training cost function

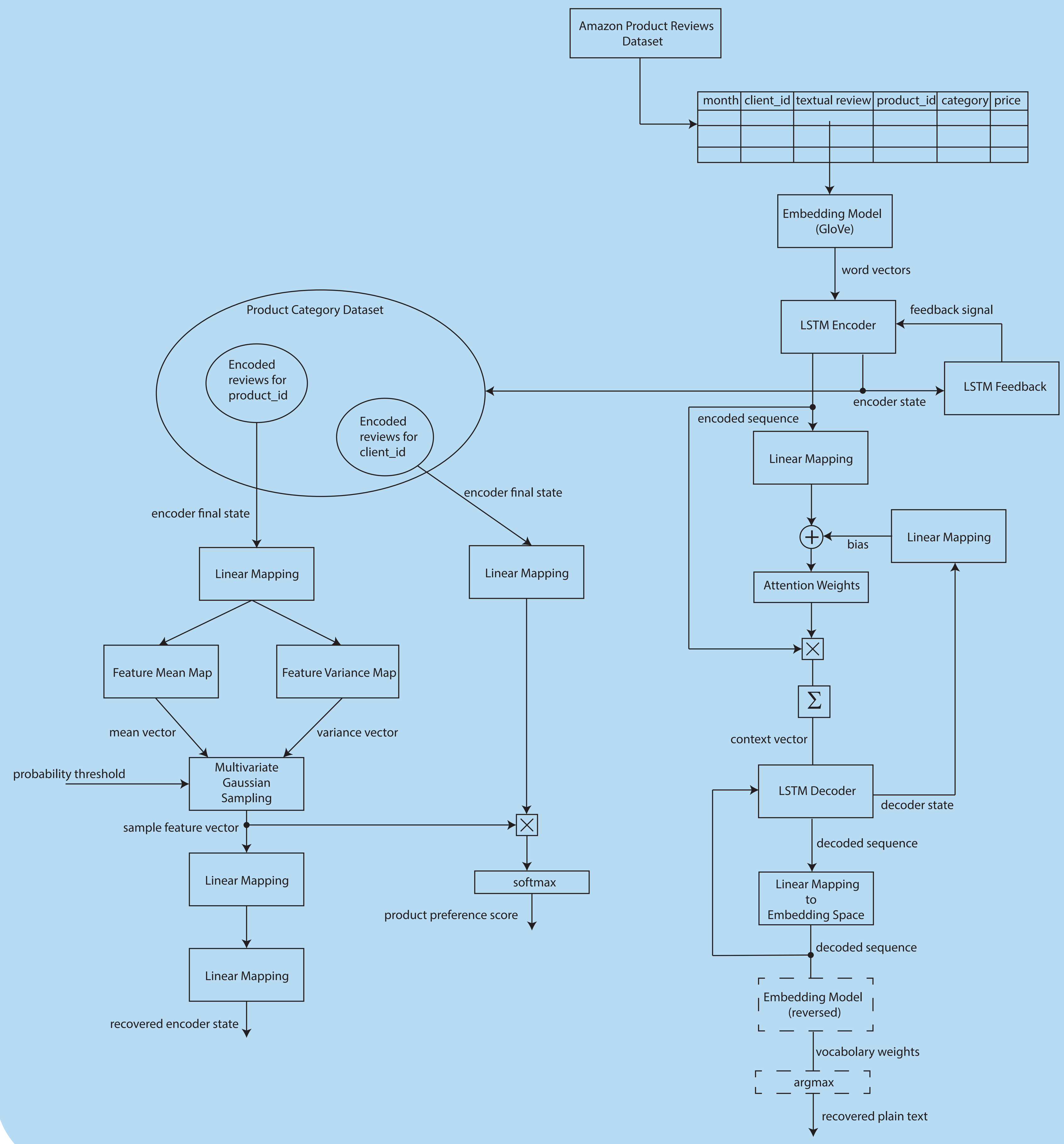
$$J_{MT} = J_{MT,reg} + J_{MT,ch} + J_{MT,vae}$$

$J_{MT,reg} = E_{X_C}(W_\mu(X_P, e_c)^2 + W_\sigma(X_P, e_c)^2).$ Regularization term for mean and variance

$J_{MT,ch} = E_\pi(-\ln(p_M)),$ Cross entropy for predicted client preference for a product

$J_{MT,vae} = E_{X_P}(e_p - \hat{e}_p),$ Difference between a input and decoded product vectors

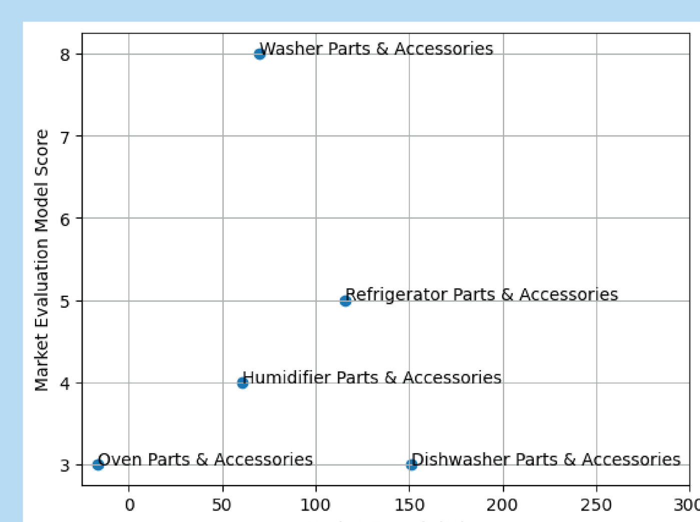
## MODEL STRUCTURE



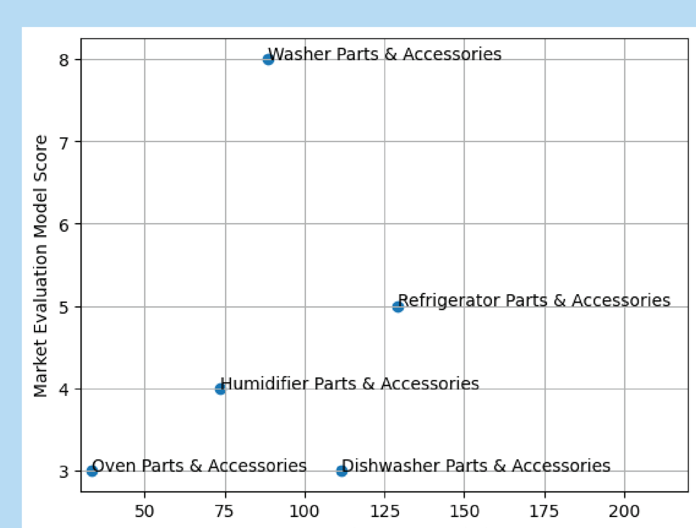

## RESULTS

The following figures present the obtained results for 5 categories of Amazon reviews. The design gap in a given market segment is predicted by looking for low probability designs through a Monte-Carlo sampling of the underlying multivariate probability distribution. The low probability products are evaluated with respect to the consumer preference probability function.

To validate the correctness of such predictions, we compare the observed market growth in the 5 segments from 2013 to 2018 year. Observe positive correlation between models scores and capitalization growth, which means that the model correctly predicts future capital allocation in the observed sectors.




Also observe the product survival rate vs market prediction, where again we see a positive correlation with the model predictions. The product survive rate is characteristic to how much a given product is in demand during the observed period.




Correlation between market scores and actually appearing new products in the observed domains is again positive where the predicted design gap is predicted. For a reference of the size of the market segments, market capitalization at the starting year is calculated.