Deep Contextual of Document Using Deep LSTM Meet Matrix Factorization to Handle Sparse Data: Proposed Model

by Mahasiswa Mahasiswa

PAPER • **OPEN ACCESS**

Deep Contextual of Document Using Deep LSTM Meet Matrix Factorization to Handle Sparse Data: Proposed Model

To cite this article: Hanafi *et al* 2020 *J. Phys.: Conf. Ser.* **1577** 012002

Deep Contextual of Document Using Deep LSTM Meet Matrix Factorization to Handle Sparse Data: Proposed Model

Hanafi[1,2], N Suryana[2] and A S H Basari[2]

[1]Department of Computer Science, Universitas Amikom Yogyakarta, Jl. Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta 55283 Indonesia.
[2]Faculty of Information and Communication Technology, University Teknikal Malaysia Malaka, Malaka 76100 Malaysia.

Abstract. Recommender system is important tool in big data era. It has responsible to make suggestion about product or service automatically for web application or mobile. In everyday utility, we cannot escape for information about food, travelling, social network, ticketing, news and etc. What the best choice for customer necessary is recommender system task to provide relevant information. Collaborative filtering is most useful recommender system technique in which considering user behaviour in the past to calculate recommendation. The first generation of collaborative filtering exploit statistical approach to calculate product recommendation. However, traditional collaborative filtering facing serious problem in scalability, accuracy and shortcoming in large data. Model based in the second generation of collaborative filtering to produce product recommendation where this model rely on matrix factorization to produce recommendation. Model based proven better performance over memory based. However, model-based performance degrades significantly when met with sparse data due the number of rating are very small. This problem popular called sparse data problem. Several methods proposed by researchers to handle sparse data problem. Mostly of them exploit text document to increase recommendation performance. However, majority of model fail to gain text document understanding. This study proceeds ongoing process with several stage. First, develop model to interpreted text document using LSTM aims to capture contextual understanding of document. Second, integrated LSTM with matrix factorization. This step aims to produce rating prediction considering text document of the product. The first step completely finished. According to experiment report, this model success to capture contextual of the document then transform into 2D space text document representation. For the further research, we are going to integrated with matrix factorization and evaluation result of rating prediction using RMSE metric evaluation.

1. Introduction

Recommender system first developed in 1990 aims to provide product information to customer specifically for e-commerce business. Recommender system is system intelligent to serve product information automatically. The characteristic information must be relevant, personalised, unique and serendipity. The most successful recommender system model is collaborative filtering due it accommodates the information characteristic mentioned on above. Even though, there are some kind of recommender system model such as content based, knowledge based, and demographic filtering based. Collaborative filtering mechanism essentially fully depend on characteristic user behaviours in previous activity. In the other hand, collaborative filtering work by calculating similarity behaviours in the past. The most popular activity in the perspective collaborative filtering is rating as representation of product or service satisfaction expression. However, number of rating collected from users are too small. So, the collaborative filtering faces the problem in minimum rating problem also popular called sparse data. The example of most popular recommender system shows on figurebelow. Content based work in practical application show on Figure 1A where they show recommendation following to product categorization while Figure 1B show the collaborative filtering mechanism.

Figure 1. Content based mechanism.

Figure 2. Collaborative filtering matrix.

Early collaborative filtering compute the recommendation following to statistical approach such as Cosine, Spearman rank, Adjust Cosine. This statistical model famous called memory based or neighbourhood model. Model based recommendation model very popular in 90s due they have simple to implemented, no require to training data. However, memory based having essential shortcoming in scalability point of view also quite high risk in growth of data. In the 2006 when Netflix competition was held, the big transformation technology of recommendation model was happened. The competition audience compete to reach best achievement in accuracy. The tremendous achievement reach by novel model dubbed model based or matrix factorization also called latent factor. Several latent factor models proven better perform over traditional memory based.

Model based fully work rely on matrix factorization approach. It can be work by rotate, reduce and completing value to finding correlation among users and products. One of the famous matrix factorization model is Singular Value Decomposition (SVD), first introduce to develop collaborative filtering by [1], enhance to increase robustness in sparse data by [2]. He also adds dimensional temporal effect. Temporal effect gained from time stamp customer take a rating. Another enhancement of SVD developed by [3]. They use probabilistic approach to advance traditional matrix factorization. To estimate missing value unrating product, they consider to applied gaussian normal distribution. That why the model dubbed probabilistic matrix factorization. Even though latent factor based on matrix factorization always stronger than memory based, it has several weaknesses. One of serious problem with latent factor is the performance recommendation degrade seriously when met sparse data (minimum rating). This research proposed novel method to eliminate sparse data on matrix factorization by integrating latent factor and document latent factor using LSTM model aims to deep understanding of the product document.

2. Previous Work

The performance of matrix factorization in sparse data condition require to increasing. The performance degrade on latent factor are very difficult to eliminate, following to Koren [4], this reasoned the missing value is too dominant. Several studies have been made by incorporating with several auxiliary information of the product. Some of them by explore image perspective

[5] to develop fashion online commerce. There are study involve audio extraction to eliminate semantic gap among text document and acoustic representation to produce online music commerce recommendation [6][7][8]. The majority study involves text document as product characteristic representation. There are several product document interpretations integrated with matrix factorization that with high impact to increasing performance in rating prediction show on Table 1 below. According author review, most of them exploit traditional NLP method to interpret the document such [9][10], and another on using modern extraction machine to interpreted product document such deep learning family include [11][12][13]. Indeed, mostly deep learning that adopted in this model leads to traditional bag of word mechanism such as auto encoder and CNN.

Table 1. Document understanding model to interpreted product review

Method	Description	Reference
Topic Modelling	Consider integrating latent factor with document context where they consider topic modelling use LDA to interpreted product document representation.	[9]
CTR	(Collaborative topic regression) They implemented topic modelling to interpreted product document. This is the method to make word classification according to similarity meaning. It also categorical LDA model.	[10]
CDL	(Collaborative deep learning) This model exploits deep learning based on auto encoder to interpreted product document meaning. This performance model better over traditional bag of word use LDA model.	[11]
HCDL	(Hybrid collaborative deep learning) This model categorical enhancement of CDL approach considering stack denoising auto encode and hybridization with SVD.	[12]
ConvMF	(Convolutional matrix factorization) This model leads to involve CNN to interpreted product document. They consider incorporating with probabilistic matrix factorisation to produce rating prediction.	[13]

3. Proposed Method
The model aims to handle inaccurate prediction obtain by matrix factorization. Our strategy considering text document owned by product. This model consist two primary work include generate deeper understanding of the document with different approach by capturing contextual use LSTM and generate rating prediction use matrix factorization. The deeper understanding method expected to improve matrix factorization against sparse data problem. The detail architecture called LSTM-PMF show on Figure 5

3.1. Architecture of LSTM and PMF
The LSTM-PMF divided six processes stage as follows: 1) Datasets collection involve two datasets from Movielens as representation explicit user expression contain user, item, rating.

2) Text document processing. This step consists several processes to transform text data into numerical mode. We applied NLTK library to conduct preprocessing application. 3) Word embedding process. Considering capturing word vector representation. Word embedding is the model to enhance word vector representation to calculate word vector 2D space dimensional. 4) Capturing contextual insight use LSTM mechanism. The benefit of LSTM training process is it have strong characteristic in sequential aspect. Sequential aspect is very important step in capturing contextual understanding. 5) Produce rating prediction use matrix factorization by considering 2D word vector representation. Aims to bridging between document vector and user vector representation involve maximum a posteriori (MAP). 6) Training and evaluation process use RMSE metrics.

Evaluation metric required to evaluate rating prediction result by comparing actual rating and output rating prediction. This research considers to implemented two datasets includes Movielens 1M (ML.1M) and 10M (ML.10M). Involvement two categories dataset consider accommodating middle population dataset and gigantic dataset population. Dataset (ML.1M) contain normal sparse data and datasets (ML.10M) contain extreme sparse data in gigantic data population. The third dataset involve movie review collecting by IMDB movie review contain 10.000 movie product. The document vector gained by LSTM will integrated with MovieLens dataset include Movielens ML.1M and Movielens ML.10M.

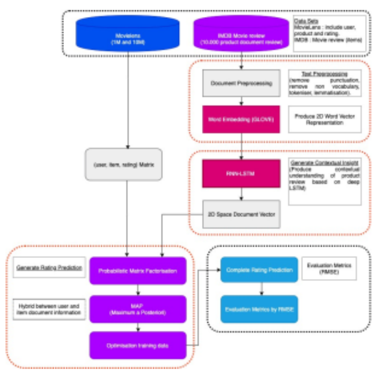

Figure 3. LSTM-PMF Architecture

3.2. Capturing context use LSTM

The basic LSTM mechanism suitable to generate sequential aspect of the sentences document. According to conventional neural language perspective, overall input is independent from another. However, in some computation point of view that is not good solution. Recurrent neural network (RNN) where LSTM is particular of RNN. Basic idea of LSTM is calculating for each network with sequence mechanism where the output gained from the previous calculation process. In the other hand, LSTM remember where it captures information correspondent to the result which have been calculate. In this framework, aims to generating document contextual insight involve 300 hidden state of LSTM for each product review. Our schema according to existing work involving document of product review. The model of LSTM to extract contextual understanding of the document can be seen on Figure 3.

Several processes required to produce word vector contextual insight representation to support LSTM model. Hybridization between pre-processing word embedding base GLOVE before transfer into LSTM input. There are several options to transform 2D dimensional vector such

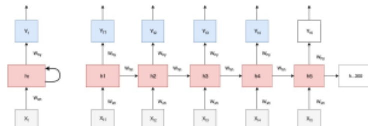

Figure 4. Design of LSTM to capture contextual insight.

as 50, 100, 300- dimension. It was needed to transform text document into 2D word vector representation. After that,processing LSTM would conduct twice to calculate and transform into contextual understanding in the form of 2D dimensional vector 50. The detail model of LSTM process shows on figure 8 includes several pre-processing.

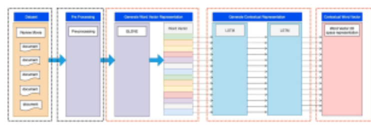

Figure 5. Document contextual extraction process

3.3. Incorporate LSTM and PMF

This model adopted matrix factorization model by Salakhutdinov [3]. This model uses probabilistic mechanism to estimate data distribution. The benefit of matrix factorization is they compatible to integrate with another data vector. Indeed, the output of LSTM cannot generate rating prediction directly. In this case, probabilistic matrix factorization become bridging between LSTM and PMF to generate rating prediction. Our proposed LSTM-PMF model show on figure 9 where red dash is representation of contextual insight and blue dash is representation of PMF field. The detail explanation of our model as follows: U represent of user vector gained Movielens datasets includes variance of user by σU 2 , R is representing of rating prediction output, W obtain document vector representation gain by IMDB.

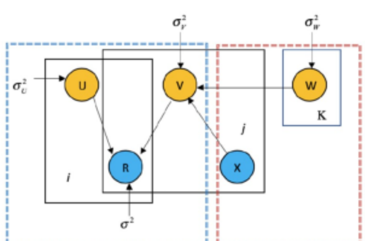

Figure 6. Data flow of LSTM and Matrix factorization

The formula rule can be explained suppose we have N users, M item, then the rating prediction can be calculated$R \in R^{NxM}$. So, the correspondent between items and users' latent factor by $U \in R^{NxM}$ and $V \in R^{NxM}$. Following to probabilistic mechanism, the observing rating formula approach as follows:

$$p(R|U, V, \sigma^2) = \Pi_i^N \Pi_j^M N(r_{ij}|u_i^t v_j, \sigma^2)^I ij \quad (1)$$

Journal of Physics: Conference Series **1577** (2020) 012002 doi:10.1088/1742-6596/1577/1/012002

Formula N x — μ, σ^2 is representation of probability density function (pdf) of the Gaussian normal distribution with mean $mu and variance σ^2, and Iij is indicator function. Aims to generate the user latentmethod, we implemented conventional prior by zero-mean spherical Gaussian prior on user latent model with variance σ_u^2. The complete formula denoted by:

$$p(U|\sigma_u^2) = \Pi_i^N N(u_i|0, \sigma_U^2 I) \tag{2}$$

3.4. Hardware and tools

The LSTM-PMF exploit large dataset and implemented sub class of deep learning machine, certainty this is required high computation. We consider utilizing GPU (graphical processor unit) and several libraries for deep learning machine embedded in Python machine such as TensorFlow, Keras and NumPy. Table 2 show the detail tools, hardware, application and library used in the LSTM-PMF experiment.

Table 2. Tools and library used

No	Device/library	Specification
1	Processor	Xeon Quad core, 2.4 GHz
2	Memory	16 Gb
3	GPU	GeForce GTX 1001
4	Tensor	Flow Deep learning tools
5	Keras	Deep learning tools
6	Anaconda	Web interface
7	Python	Tool programming
8	Scikit-learn	Handle sparse module
9	NumPy	Matrix Factorizations
10	NLTK	NLP tools

4. Text Document Pre-processing

The first process step requires to conduct pre-processing aims to transform text to fulfill NLP standard. And also, text document information cannot directly to compute by LSTM. This research considers to utilized NLTK module that available on Python standard module also can be downloaded for free. Several important in pre-processing stage include tokenizer, removing stop word, remove punctuation, remove non vocabulary English verb and lemmatization. Detail table 3 on below explain several preprocessing steps.

4.1. Result and analysis

This study consists three particular step including GLOVE word pre-processing, generate 2D document latent vector and integrating document latent factor vector into matrix factorization.

Table 3. Document pre-processing scenario

No	Text Pre-processing step
1	Setting maximum length of sentences up to 300 word, this step aims to capture of the contextual meaning of the document. The limiting role taken according previous study decided with 300 maximum word.
2	Delete stop words in the sentences, this step aims delete unused characters and symbol which it does not influent the meaning.
3	Calculate TF-IDF score for each document, the objective is deleting the score of words that too much dominant in every document review on pre-processing datasets.
4	Remove all of non-vocabulary words from catalogue sentences, this step are not influent the meaning context and it can reduce the computation cost.

At present, this experiment result achieves second process in generating 2D space document latent factor use GLOVE LSTM. We possibly evaluate and comparing our result of proposed method with existing previous work after our experiment completely finished.

Table 4: Output of LSTM document vector 2D space

No.	Text document of product review	2D document vector representation
1.	Sentence: snobbish professor agrees wager can make flower girl high professor so sure abilities takes it upon himself transform girl into someone can pass cultured member high subject turns out be lovely agrees speech lessons improve her job then form unlikely bond one threatened aristocratic witty adaptation musical about takes bet can transform dirty flower girl into fool everyone into thinking she really thus young aristocrat falls madly love takes all credit forgets acknowledge her leaves him suddenly realizes grown her face really live without chance meeting between two leads wager will test they hear flower girl proposes transform into lady aristocratic some agrees become their test	Last RNN State: [0.00361181,- 0.00629696, 0.00137311, 0.00107181, 0.00545287,-0.00289809, - 0.00465942,-0.00034414,-0.00564274, 0.00163962,- 0.00050874,-0.00306838, 0.00070256, 0.00073792, 0.00111825, 0.00357915, 0.0055963 , 0.00048963, 0.00142733, 0.00410594, 0.00392801,- 0.00319667,- 0.00116569, 0.00031913, -0.0066682 , 0.00465886, 0.00421595,- 0.00659875, 0.0038083 ,-0.00079539, -0.00293556,-0.00198917,-0.00364866,- 0.007879 ,- 0.00184936,-0.00879543, 0.00419056, 0.00134126,- 0.00361824,- 0.00650992, 0.00085606,-0.00057412, 0.00248335,-0.00092162,-0.0023925 ,- 0.00089824,- 0.00294857, 0.00063654, 0.00484379,-0.00152714,- 0.00143955,- 0.00127011, 0.00571889,-0.00030334, 0.00119753, 0.00097439,-0.00088358,- 0.00251328,- 0.00381235, 0.0038111 , -0.00238481,-0.00252997, 0.00376203, 0.00265672, 0.00167894,-0.00457901, - 0.00305263, 0.00467645, 0.0023127 , 0.00264076, 0.00631951, 0.00161645, -0.00298681, 0.00462157, 0.0038949 ,- 0.00039923,-0.00015931,-0.00091718, 0.00347933,-0.00594147, 0.00434831,- 0.0046596 ,- 0.00297752,-0.00020768, -0.00540692, 0.00658523,- 0.00292715, 0.0017341 , 0.0003999 ,-0.00172803, 0.00135662, 0.00123143,-0.00266397, 0.00446418,- 0.00388802, 0.00105405, 0.00060515,-0.00036911,- 0.0000576 , 0.00143934]

2	Sentence: unemployed single mother becomes legal assistant almost brings down power company accused water unemployed single desperate find having no losing streak even failed against doctor car accident she was no she successfully her lawyer give her job compensation no one takes her her clothes soon changes she begins investigate suspicious real estate case involving she discovers company trying quietly buy land was deadly toxic waste company dumping residents she digs finds herself leading point series events would involve her law firm one biggest class action history against dollar woman tight car accident which not at pleads her attorney hire her at law stumbles upon some medical records placed real estate convinces allow her where she discovers involving water local community which causing devastating among its stars this legal drama based story woman helped win largest settlement ever paid single mother three after losing personal	Last RNN State: [-0.00058141, 0.00385763, 0.00123018, 0.00427003,- 0.00048762,-0.00658738, -0.00196021,-0.00348914,-0.00326066,- 0.00253979,-0.0028013 , 0.0001128 , 0.00302504, 0.00011595, 0.00053489, 0.00100057,-0.00119514, 0.00301734, -0.00130226,-0.0037403 , 0.00736278,- 0.00387619, 0.00695837,-0.00379596, 0.00194974,-0.00117637,-0.00741686, 0.00372151,-0.00384646, 0.0034766 , 0.0007024 ,-0.0042891 , 0.00371806, 0.00556516,-0.00111934,-0.00188738, -0.00103761,-0.00123387, 0.0044857 , 0.00417872, 0.00407807,-0.00708222, -0.0023213 , 0.00201266,-0.00164627,- 0.0048811 ,-0.00269887,-0.00013715, -0.00076007,-0.0053246 ,-0.00139777, 0.00401163,-0.00141868,-0.00589907, 0.00180107,-0.00460449,-0.0012688 , 0.00108053, 0.00454185, 0.0036293 , -0.00149622, 0.00096825, 0.00041592,- 0.00036147,-0.00327778,-0.00010784, -0.00288902,-0.00305635, 0.00176236, 0.00252666,-0.00359611, 0.00103384, -0.00549097, 0.00519081,-0.00023094,- 0.00044565, 0.00399532,-0.00461778, 0.00140178,-0.00785419, 0.00334468, 0.00298712,-0.00048999, 0.00187281, -0.00044592,-0.00147288,-0.00298558,- 0.00401561,-0.00147517,-0.00154868, -0.00609612, 0.00553525, 0.00548647,- 0.00146683, 0.00467391,-0.00251919, 0.00851225, 0.00373096, 0.00105892, 0.00406634]

Journal of Physics: Conference Series **1577** (2020) 012002 doi:10.1088/1742-6596/1577/1/012002

3.	Sentence: misfit looking save colony greedy recruits group bugs turn out be inept circus annual huge colony ants forced collect every piece food grows their island group all changes misfit inventor ant named accidentally over offering pile thus forcing devious leader force ants their fact friends believe him desperate help save volunteers go out into world search group what got was talented group circus grasshoppers return take control must prove himself true hero before too every greedy grasshoppers demand share food colony one their annual offering inadvertently grasshoppers demand twice much food ants send misfit ant named seek battle only discover group bugs recruited are inept circus bugs are only hope fight off their good her are absolutely against annual demands this things are bound get bad embarrassing terrible threatens already balance between two only resourceful young ant	Last RNN State: [0.00081602,- 0.00241877,- 0.0049815 , 0.00301353, 0.00386939, 0.00471222, 0.00455486,-0.00160164,-0.00062518, 0.00046425, 0.00509222, 0.00452081, -0.00766711, 0.00933722,- 0.0024412 ,- 0.00159593, 0.00197213,-0.00163494, -0.0011714 ,-0.0003353 ,-0.00367733, 0.00003134,- 0.00219219, 0.00535094, 0.00340234, 0.00386255, 0.00416878,- 0.00223821, 0.00220804, 0.00045752, 0.00326856, 0.00342699,-0.00170999,- 0.00010439,- 0.00269502,-0.00145216, 0.00112151,-0.00086121, 0.00011781, 0.0037838 , 0.00603624,-0.00452917, 0.00242836, 0.00541125, 0.000262 , 0.00727725, 0.00099493, 0.00265136, -0.00053416,-0.00348472,- 0.00043007,- 0.00111197, 0.00175155, 0.00284908, -0.00037088,-0.00940969, 0.00385419, 0.00440253, 0.00169279,-0.00063607, -0.00048265, 0.00011516,- 0.00312393, 0.00214836, 0.0000828 ,-0.0026947 , - 0.00478126, 0.00761514, 0.00334759,- 0.00298276,- 0.00570504, 0.00107021, 0.00427708, 0.00395506, 0.00119789, 0.00230227, 0.00441155, 0.00037845, - 0.00310474,-0.00204851, 0.00373436, 0.00149822,- 0.0012248 , 0.00094843, 0.00170021,-0.00157434, 0.00018996,- 0.00016627,-0.00229208,-0.00465714, - 0.00273689,-0.00686116, 0.00001319, 0.00457163, 0.0038427 ,-0.00103517, -0.00121602,-0.00333051, 0.00171823, 0.00219404]

5. Conclusion

According to experiment result on pre-processing and LSTM step, the model successful to generate document latent representation in 2D space 100 vector use IMDB movie review including 10.000 document product. In further research, we are going to training in datasets ML-1M and ML- 10M expected to increase performance of matrix factorization in rating prediction. Contextual insight expect increase correspondent between product and user effectively due contextual insight will produce high score for contextual document representation and low score in non-contextual representation. Hight score certainty will influenced high correspondent between item and product.

6. References

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Application of dimensionality reduction in recommender system-a case study. Technical report, Minnesota Univ Minneapolis Dept of Computer Science, 2000.

[2] Koren Y Collaborative. filtering with temporal dynamics [j]. *Communications of the ACM*, 53(4):89–97, 2010.

[3] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[4] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[5] Shatha Jaradat. Deep cross-domain fashion recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 407–410, 2017.

[6] Qing Li, Byeong Man Kim, Dong Hai Guan, and Duk whan Oh. A music recommender based on audio features. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 532–533, 2004.

[7] Kibeom Lee and Kyogu Lee. Using dynamically promoted experts for music recommendation. *IEEE transactions on multimedia*, 16(5):1201–1210, 2014.

[8] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.

[9] Guang Ling, Michael R Lyu, and Irwin King. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 105–112, 2014.

[10] Xiaolin Zheng, Weifeng Ding, Jingnan Xu, and Deren Chen. Personalized recommendation based on review topics. *Service Oriented Computing and Applications*, 8(1):15–31, 2014.

[11] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1235–1244, 2015.

[12] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. Collaborative filtering and deep learning based hybrid recommendation for cold start problem. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 874–877. IEEE, 2016.

[13] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 233–240, 2016.

Acknowledgments

We would to thank for institution who always supporting us to funding our research. Our great thank Time Excelindo company, Universitas Amikom Yogyakarta and University Teknikal Malaysia Malaka (UTeM).

Deep Contextual of Document Using Deep LSTM Meet Matrix Factorization to Handle Sparse Data: Proposed Model

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12