JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 21, No. 5, pp. 951-958, May 2020

Check for updates

루간다어 감성 분류를 위한 저자원 유튜브 댓글 인코딩

Abdul Male Ssentumbwe[1] · 정 유 철[2] · 이 현 아[3*] · 김 병 만[3]
[1]마케레레 대학교 비즈니스 스쿨, 컴퓨터과학 및 공학과 강사
[2]금오공과대학교 컴퓨터공학과 교수
[3]금오공과대학교 컴퓨터소프트웨어공학과 교수

Low-resource YouTube comment encoding for Luganda sentiment classification performance

Abdul Male Ssentumbwe[1] · YuChul Jung[2] · Hyunah Lee[3*] · Byeong Man Kim[3]

[1]Lecturer, Department of Computer Science and Engineering, Makerere University Business School, Plot 21 A, Port Bell Rd, Kampala, Uganda
[2]Professor, Department of Computer Engineering, Kumoh National Institute of Technology, Gumi, 39177, Korea
[3]Professor, Department of Computer Software Engineering, Kumoh National Institute of Technology, Gumi, 39177, Korea

[요 약]

최근 소셜네트워크의 발달로 저자원언어의 다중언어 의견데이터도 증가하고 있다. 우간다에서 사용되는 루간다어(Luganda)는 저자원언어로 유튜브에서의 감성 분석을 위한 루간다 코퍼스를 획득하기 쉽지 않다. 본 논문에서는 유튜브 동영상의 루간다어 의견 감성 분석을 위한 댓글 수집 방안을 제안하고, 선택된 기계학습과 딥러닝 분류 알고리즘을 사용하여 수집된 158개의 댓글의 적합성을 평가한다. 주어진 저자원 상황의 루간다어 댓글에 대한 10겹 교차검증에서, 수집된 데이터는 기계학습에서는 Gaussian Naive Bayes(55%), 딥러닝에서는 Multilayer Perceptron sequential model scoring (68.8%)이 가장 좋은 성능을 보였다.

[Abstract]

The recent boom in social networks usage has generated some multilingual opinion data for low-resource languages. Luganda is one of the major languages in Uganda, thus it is a low-resource language and Luganda corpora for sentiment analysis especially for YouTube is not easily available. In this paper, we propose assumptions to guide collection of Luganda comments using Luganda YouTube video opinions for sentiment analysis. We evaluate the suitability of our clean YouTube comments (158) dataset for sentiment analysis using selected machine learning and deep learning classification algorithms. Given the low-resource setting, the dataset performs best with Gaussian Naive Bayes for machine learning (55%) and deep learning Multilayer Perceptron sequential model scoring (68.8%) when dataset splitting is at 10% for test set with Luganda comment segmentation.

색인어 : Luganda, 저자원 언어, 감성분석, 유튜브 댓글, 의견 마이닝

Key word : Luganda, Low-resource language, Sentiment Analysis, YouTube Comments, Opinion Mining

http://dx.doi.org/10.9728/dcs.2020.21.5.951

Received 04 March 2020; **Revised** 15 May 2020
Accepted 25 May 2020

***Corresponding Author; Hyunah Lee**

Tel: +82-54-478-7546
E-mail: halee@kumoh.ac.kr

Ⅰ. Introduction

Luganda or Ganda is one of the major languages in Uganda (a region in East Africa) used by more than 6 million people from the central region. The language is mainly used by Baganda people from Buganda tribe, and belong to the Bantu language family [1]. However, English and Swahili are the main official languages as per the constitution [2] and the country is blessed with about 43 as many languages in total according to Ethnologue Languages of the World [3]. With all that said, Luganda is the second widely used language after English and thus the core motivating factor for our interest into sentiment analysis for YouTube Luganda comments.

Luganda YouTube comments, opinions, or reviews are generally not easy to find separately in existence, first because mostly online users comment in a mix of both English and Luganda or with another local language. This is a result of English usage for education throughout years by natives and sometimes comments are mixed with Arabic transliterations due to borrowed words from Muslim faith. Secondly, because Luganda is a low-resource language and Luganda corpora for sentiment analysis especially for YouTube is not easily available. With such scarcity in Luganda datasets, for one to conduct sentiment analysis especially targeting low-resource languages can be a daunting task.

Recently, due to increased access to internet, Uganda online user numbers have also grown [4]. This applies that online data such us from social media platforms for local low-resource languages is likely to grow with time. As we all know, technology keeps changing quickly and researchers especially for low-resource languages need to use every opportunity to test new technologies otherwise low-resource communities may lag behind in acquiring skillset due dataset unavailability.

Sentiment analysis is a way for summarizing and determining people's feelings or opinions based on their comments or reviews gathered from online platforms such as Social Media sites (such as YouTube, Facebook, Twitter etc.), ecommerce sites (such as Amazon, Alibaba etc.), blogs and news sites [5]- [7]. A number of researchers have used YouTube data for mining online user opinions or sentiments. Choudhury and Breslin [8] based on top 4000 videos from different 5 classification YouTube categories, proposed unsupervised lexicon technique for detecting sentiment polarity in video comments from users. They reported better results in detecting user sentiments with the idea of extending SentiWordnet with their lexicon technique. Similarly, a recommender system based on analysis of people's YouTube channel comments was designed for detecting emotion sentiments. The detected sentiment then is used as a basis for recommending appropriate media items to users [9].

A part from YouTube, most works on sentiment analysis gained a lot popularity with increased use of Facebook and Twitter platforms by various internet users. For example, some authors based on Myamar movie comments from Facebook designed a system for automatically assigning polarity scores [10]. In Uganda, sentiment analysis has generally focused more on Facebook and Twitter. Using twitter comments from Uganda traditional media houses, some authors report that generally user sentiments related to politics, security and economy topics were found to be more negative and sports comments tended to be more positive [11]. In addition, other researchers based on #UgandaDecides, conducted a voter sentiment analysis for the 2016 Uganda presidential elections from comments generated on Twitter. The authors found a challenge in data collection since the targeted English comments had a mix of local languages, which ultimately were dropped [12].

Against this background, gathering enough data is normally not easy in an environment of low-resource context and requires a set of guidelines to acquire appropriate data for research [13], [14]. In this work, we propose some assumptions in order to be able to gather useful Luganda YouTube comments for sentiment analysis. We focus on Luganda comment gathering and we perform various evaluation to test the usefulness of the collected comments in supporting experiments for sentiment analysis in a low-resource environment.

Ⅱ. Proposed Approach

2-1 Determining Luganda YouTube comment collection

There exists many YouTube music channels for/from Ugandan audiences. These channels tend to have generally positive emotion or opinions and our study requires a scope, which is representative of both positive (pos), negative (neg) as well as neutral (neu) comments. Similarly, Ugandan artist channels in existence largely suffer from the same scenario above. This is so because the channels tend to align more to the artist's supporters or fun base. Others like disco jokers (DJs) and music promoter channels etc. we find that generally comments there also tend to be positively inclined. In view of the above, we propose Ugandan channels criticizing artists' music videos. These tend to carry relatively well-balanced emotionally oriented comments appropriate for this study.

Our desire is to harness as many Luganda comments as

possible but we have to remember that comments from these channels are mixed i.e. English and Luganda in one sentence and sometimes with possibility of finding additional local languages in use too. In such circumstances, there is need to come up with assumptions for sorting or filtering and finally selecting Luganda comments to building our dataset for sentiment analysis modelling.

Assumptions for Luganda comments collection

- Identify appropriate YouTube channel criticizing artists' music videos.
- Identify if a channel's given video has more than 5,000 views with at least 5 comments.
- Use channel YouTube video link to download all comments.
- Remove wholly English sentence comments since our target is Luganda comments.
- Keep a mixed sentence comment if it has more than 2 Luganda words.

2-2 Luganda YouTube dataset

We follow the proposed guidelines to identify our channel for target Luganda comments, and we identified a TV show called "*Akasengejja*" broadcasted on Bukedde TV a local channel in Uganda [15]. The station broadcasts mostly in Luganda and the TV show chosen for our comment gathering focuses on assessing Ugandan artists' music videos where later the aired show is posted on their Akasengejja YouTube channel. Using a YouTube comment scraper [16], which is based on Google YouTube API [17], we download 294 mixed comments using 18 YouTube video links from our chosen channel satisfying our assumptions. Fig 1. shows an example of Luganda comments in YouTube. After, we manually remove wholly English comments and finally left with 158 mostly Luganda comments as per set guidelines. Since one of the authors is a native Luganda speaker, to reduce subjectivity during comment label assignment or annotation, we adopt one additional native through online collaboration and together we assign each of the 158 Luganda comments with a general sentiment label. A comment is read and determination is made as to whether it's a positive (1), negative (-1) or neural (0). The Table 1, shows the total number of comments after general sentiment assignment (label) in our Luganda dataset and in Table 2, we also show a sample of 3 comments from our Luganda dataset.

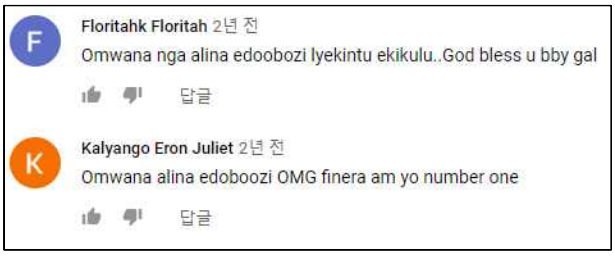

그림 1. 루간다어 유튜브 댓글 예
Fig. 1. Some example of Luganda YouTube Comment

표 1. 루간다어 댓글 데이터집합 통계
Table 1. Total number of Luganda dataset comments

Positive (1)	Negative (-1)	Neutral (0)	Total
71	48	39	158

표 2. 루간다 유튜브 댓글의 레이블링 예시
Table 2. Some example of labelled Luganda YouTube comments

Luganda YouTube Comment	Label
Luganda: *Juliana yasinga namwe kirabwa nomuto* Literal meaning: Please Juliana is the best, even a young child can notice this.	1
Luganda: *wabula naye nakoowa bi dance byabasezzi o* Literal Meaning: But I am really tired of night-dancer styles. Oh!	-1
Luganda: *Wabula omusajja batte* Literal meaning: But this man Batte.	0

2-3 Proposed token processing and normalization

Our aim is to utilize the collected Luganda comments in order to determine the most appropriate technique(s) to build a sentiment model while working in a low-resource setting. Firstly, Luganda comment data under goes through various preprocessing to make it fit for intended training experiments at each stage. For example, we tokenize and lowercase all comments, remove all punctuation marks and non-alphabetic characters. After this process we are left with 964 clean tokens and 641 unique vocabulary. During cleaning we do not use any stopword list since at the time of this writing, no known agreed upon list for Luganda language could be found.

Secondly, we normalize tokens in aspects of artists' most common name in comments, repeating words like *"hahahahaha"* are shortened to common phrase, names misspelt are also

표 3. 토큰 정규화 추출 예시
Table 3. Some example of extracted Token normalization

Token Before	Token After
Haha, Hahahahaha, haahaa	haha
hmmhmmmm, hmmmm	hmm
wawawa, waaawaa	wawa
Batte, Bate, Batey	batte
Bobiwine, Bobi wine, **Bobi**, Bobie	bobi
Remah, **Rema**	rema

Bold tokens were the most common version of an artist name used in Luganda comments and therefore used for normalization.

normalized to one common name style in comments. After normalization process, 961 clean tokens and 622 unique vocabulary is left. Table 3 shows some example of comment token normalization.

Thirdly, because sentiment analysis cannot be conducted directly on text tokens using machine learning or deep learning techniques, we transform our comments using selected text encoding techniques based on Bag of Word (BoW) model to enable comparison analysis. The main text encoding used include, binary, count, tf-idf and frequency representation or feature extraction methods.

2-4 Proposed transformations with comment segmentation (GandaKIT)

In a recent machine translation (MT) research for English to Luganda, the authors designed GandaKIT as a tool for morphological segmentation of Luganda sentences [18]. As a result of their tool, they demonstrated improved performance in MT output using Statistical Machine Translation with Moses toolkit.

In similar way, we propose to adopt GandaKIT segmentation function to our normalized Luganda YouTube comments for further comparison of comment sentiment analyses. We believe the additional process here too, is capable of improving our classifier training models and subsequent prediction on tests.

Ⅲ. Experimental Process

3-1 Dataset

As in Table 1, the dataset is composed of 158 labeled Luganda comments with positive (1), negative (-1) and neural (0) labels. We conduct experiments using normalized clean comments after all preprocessing steps described before.

3-2 Classifier modelling

- **Machine Learning Model (MLM)**: From sklearn machine learning library, we utilize Gaussian Naive Bayes (GNB), LinearSVC from Support Vector Machines (SVM) and Random Forest (RFC) classifiers to train encoded vector comments. Then we evaluate each classifier model's prediction accuracy for test set based on comparison between text encoding technique used (i.e. binary, count, tf-idf, and frequency). The chosen classifiers inherently can support multi-class classification which is the case for our sentiment problem at hand.
- **Neural Network Model (NNM)**: From Keras deep learning library, we utilize Multilayer Perceptron (MLP) for multi-class softmax classification based on sequential model. The input layer for this model is guided by the number of words (622 normalized comments), one hidden layer shaped at 100 with 'relu' as activation function and final output layer shaped at 3 because of 3 classes in our sentiment prediction with 'softmax' as activation function. We compile our NNM model with 'categorical_crossentropy' as loss function, optimized with 'adagrad' and 'accuracy' for generating metric scores. The model is also trained at 20 epoch and evaluated for test set split at 10%, 20%, and 30% for each encoding. The whole process is automatically repeated 10 times where we report the average scores for each encoding method.

Ⅳ. Results and Discussion

4-1 MLM Results with cross validation

We use 10 fold cross validation and report mean accuracy scores for each machine learning classifier based on encoding technique used. In Table 4, evaluation results based on test set splitting at 10%, 20% and 30% with cross validation is presented.

At a glance, we can notice that at 10% test set split, the GNB classifier performed generally better and thus leading in the 4 encoding methods. This could probably be that GNB benefits

표 4. CV 10%, 20%, 30% 데이터인코딩별 MLM 분류 평균 점수
Table 4. MLM Classifier mean score (%) per encoding on Test set at 10%, 20% and 30% with CV

Classifier Model	Text encoding method (Test set 10%)			
	Frequency	Binary	Count	Tf-Idf
GNB	**50.0**	**50.0**	**50.0**	**45.0**
SVM	35.0	35.0	35.0	40.0
RFC	20.0	30.0	40.0	25.0
Classifier Model	**Text encoding method (Test set 20%)**			
	Frequency	**Binary**	**Count**	**Tf-Idf**
GNB	34.2	34.2	34.2	37.5
SVM	**50.8**	**41.7**	41.7	44.2
RFC	41.7	40.0	**42.5**	**47.5**
Classifier Model	**Text encoding method (Test set 30%)**			
	Frequency	**Binary**	**Count**	**Tf-Idf**
GNB	**41.5**	41.5	41.5	**46.0**
SVM	38.0	**45.5**	**47.5**	38.0
RFC	31.5	41.0	35.0	37.5

GNB - Gaussian Naive Bayes, SVM – Support Vector Machine (LinearSVC), RFC - Random Forest classifier

from having more training data (90%). At 20% split, we can see a general reduction in GNB classifier's sentiment prediction performance though at 30% split it tries to gain a little ground. SVM notably gains ground in test set splitting at 20% and 30% exhibiting its highest score of 50.8 using frequency encoding. For the RFC classifier, it scored better at 20% test set split with encoding techniques count (42.5) and tf-idf (47.5) taking lead but registering poor results for 10% and 30% test set splits. We can notice some general performance instability in classifier scores when we increase test set split to 20% and 30%. This behaviour is expected for low-resource environment like ours because of decrease in number of training samples as more samples are availed for the test set.

In Table 5, MLM classifier with comment segmentation shows a consistent improvement in performance especially for Test split at 10%. Under this, the GNB classifier reported a 5.0 overall best increase in results across the four text encoding methods.

표 5. 분할된 댓글 CV 10%, 20%, 30% 데이터인코딩별 MLM 분류 평균 점수

Table 5. MLM Classifier mean score (%) per encoding on Test set at 10%, 20% and 30% with CV with comment segmentation

Classifier Model	Text encoding method (Test set 10%)			
	Frequency	**Binary**	**Count**	**Tf-Idf**
GNB	**55.0**	**55.0**	**55.0**	**50.0**
SVM	35.0	30.0	30.0	35.0
RFC	35.0	25.0	35.0	30.0
Classifier Model	**Text encoding method (Test set 20%)**			
	Frequency	**Binary**	**Count**	**Tf-Idf**
GNB	35	35.0	35.0	31.7
SVM	**46.7**	**50.0**	**50.0**	**43.3**
RFC	44.2	38.3	47.5	40.8
Classifier Model	**Text encoding method (Test set 30%)**			
	Frequency	**Binary**	**Count**	**Tf-Idf**
GNB	48.0	48.0	48.0	**50.0**
SVM	**51.0**	**50.0**	**50.0**	47.0
RFC	46.0	37.5	29.0	37.0

GNB - Gaussian Naive Bayes, SVM – Support Vector Machine (LinearSVC), RFC - Random Forest classifier

4-2 Neural Network Model (NNM) Results

Having trained our Multilayer Perceptron (MLP) based on sequential model, Table 6 shows results for each encoding technique evaluated for test set split at 10%, 20%, and 30%.

From Table 5 our neural nework MLP classifier model on test set split at 10% scored highest (average score 66.3 for 'Tf-Idf' text encoding) followed by 59.4 for 'Count' text encoding method. However, classifier performance scores drop for test set split at 20% registering 56.9 as the highest average score based on 'Count' text encoding. The same trend is observed for test set split at 30% though the average scores drop below the 50% mark for all text encoding techniques. Similar to the MLM classifier, our MLP of NNM also suffers in performance as more samples are retained for the test set at 20% and 30% split. In addition it is known that deep learning algorithms heavily require an appropriate volume of training data in order to predict well on a test set and therefore a reduction in training data in this case does not benefit the model further at all.

표 6. 10%, 20%, 30% 데이터인코딩별 MLP 분류 비교차검증 평균 점수

Table 6. MLP Classifier mean score (%) per encoding on Test set at 10%, 20% and 30% without cross validation

Split %	Text encoding method			
	Frequency	**Binary**	**Count**	**Tf-Idf**
Test set 10%	53.4	56.3	59.4	**66.3**
Test set 20%	51.6	55.3	**56.9**	55.0
Test set 30%	47.1	48.1	47.8	**48.8**

After training the MLP Classifier model on the same dataset with comments segmented using the GandaKIT tool [17], the model results in Table 7 show an improvement in general classification performance especially for Test set split at 10%. Here Tf-Idf scored 68.8, representing a 2.5 increase in comparison to Table 6 results without comment segmentation. The reported increment in performance in segmented models is because the segmentation process assists to increase Luganda word token count which contributes to more availed data for model training thus better performance.

표 7. 분할된 댓글 10%, 20%, 30% 데이터인코딩별 MLP 분류 비교차검증 평균 점수

Table 7. MLP Classifier mean score (%) per encoding on Test set at 10%, 20% and 30% with comment segmentation without cross validation

Split %	Text encoding method			
	Frequency	**Binary**	**Count**	**Tf-Idf**
Test set 10%	**53.0**	**63.4**	**63.6**	**68.8**
Test set 20%	50.1	52.3	51.2	52.9
Test set 30%	46.9	48.9	49.1	50.3

Ⅴ. Conclusion

We proposed some assumptions for Luganda YouTube comment gathering from channels critiquing Ugandan music videos and from the experimental results we have demonstrated with our created low-resource dataset that it is essential to monitor split percentages between test and training sets in order to achieve some substantial performance especially in low-resource environment like for Luganda language. Results were presented using different text encoding methods based on BoW model to train and evaluate machine learning and neural network model classifiers. Here the neural network MLP classifier model achieved overall best average score at 66.3% using 'Tf-Idf' text encoding at 10% test set split without segmentation and the same classifier achieves the overall best score of 68.8% with Luganda comments segmented all without cross validation. In general, the ideas and result findings can be useful for future research as baselines in area of Luganda sentiment analysis for YouTube opinion mining as more comment or opinion data for Luganda language becomes readily available online.

Acknowledgment

본 연구는 금오공과대학교 교수연구년제에 의하여 연구된 실적물

References

[1] Wikipedia contributors. Luganda. *Wikipedia, The Free Encyclopedia* [Internet]. Available: https://en.wikipedia.org/w/index.php?title=Luganda&oldid=857068757.

[2] B. E. Sawe, What Languages Are Spoken in Uganda? - WorldAtlas.Com [Internet]. Available: https://www.worldatlas.com/articles/what-languages-are-spoken-in-uganda.html.

[3] D. M. Eberhard, G. F. Simon, and C. D. Fennig, Ethnologue: Languages of the World. Twenty-Second Edition. Dallas, Texas: SIL International [Internet]. Available: https://www.ethnologue.com/country/ug.

[4] Uganda Communications Commision. *Post, Broadcasting and Telecommunications Market & Industry Q3 Report, July-September 2017*. Kampala, Uganda, 2017.

[5] A. U. R. Khan, M. Khan, and M. B. Khan, Naïve Multi-Label Classification of YouTube Comments Using Comparative Opinion Mining. *Procedia Computer Science*, Vol. 82, pp. 57–64, 2016.

[6] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi, "Sentiment Analysis on YouTube: A Brief Survey." *MAGNT Research Report*, Vol. 3, No. 1, pp. 1250–1257, 2015.

[7] J. Lee, W Lee, J. Park, and J. Choi, The Blog Polarity Classification Technique using Opinion Mining, *Journal of Digital Contents Society*, Vol. 15, Issue 4, pp. 559-568, 2014.

[8] S. Choudhury, and J. G. Breslin, User Sentiment Detection: A YouTube Use Case. in *The 21st National Conference on Artificial Intelligence and Cognitive Science* (AICS 2010), 2010.

[9] E. Mulholland, P. Mc Kevitt, T. Lunney, and K. Schneider, Analysing Emotional Sentiment in People's YouTube Channel Comments. In *ArtsIT/DLI*, pp. 181–188, 2017.

[10] W. W. Thant, N. Thwet, T. Aung, S. S. Htay, K. K. Htwe, and K. T. Yar, "Assigning Polarity Scores to Facebook Myanmar Movie Comments." *International Journal of Computer Applications*, Vol. 177, No. 6, pp. 0975–8887, 2017.

[11] F. Namugera, R. Wesonga, and P. Jehopio, "Text Mining and Determinants of Sentiments: Twitter Social Media Usage by Traditional Media Houses in Uganda." *Computational Social Networks*, Vol. 6, No. 1, p. 3, 2019.

[12] I. Mukonyezi, C. Babirye, and E. Mwebaze, "Mining Voter Sentiments from Twitter Data for the 2016 Uganda Presidential Elections Claire Babirye Ernest Mwebaze." *International Journal of Technology and Management*, Vol. 3, No. 2, pp. 1–12, 2018.

[13] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual Sentiment Analysis: From Formal to Informal and Scarce Resource Languages." *Artificial Intelligence Review*, Vol. 48, No. 4, pp. 499–527, 2017.

[15] T. A. Le, D. Moeljadi, Y. Miura, and T. Ohkuma, Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. 2016. in *Proceedings of the 12th Workshop on Asian Language Resources* (ALR12), pp. 123–131, 2016.

[16] Akasengejja | Bukedde TV [Internet]. Avaialble: https://bukeddetv.co.ug/akasengejja-2/.

[15] P. Klostermann, GitHub-philbot9/youtube-comment-

scraper: A web client that scrapes YouTube comments [Internet]. Available: https://github.com/philbot9/youtube-comment-scraper.

[17] Google Inc. API Reference | YouTube Data API | Google Developers. *Google Developers* [Internet]. Available: https://developers.google.com/youtube/v3/docs/.

[18] A. M. Ssentumbwe, B. Kim, and H. Lee, "English to Luganda SMT: Ganda Noun Class Prefix Segmentation for Enriched Machine Translation," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 5, pp. 1861-1868, Oct. 2019.

Abdul Male Ssentumbwe

2004년 : Makerere University (학사)
2010년 : Sikkim Manipal University (석사)
2016년~2020년 : 금오공과대학교 (박사)

2020년~현 재 : Makerere University Business School, Uganda (강사)
※관심분야 : 인공지능, 기계학습, 정보검색, 자연어처리 등

정유철 (Yuchul Jung)

2003년 : 아주대학교 정보 및 컴퓨터공학과 (학사)
2005년 : KAIST 정보통신공학 (석사)
2011년 : KAIST 전산학과 (박사)
2009년~2013년: 한국전자통신연구원 선임연구원
2013년~2017년: 한국과학기술정보연구원 선임연구원

2017년~현 재: 금오공과대학교 컴퓨터공학과 조교수
※관심분야 : 인공지능, 기계학습, 정보검색, 자연어처리 등

이현아 (Hyunah Lee)

1996년 : 연세대학교 컴퓨터과학과 (학사)
1998년 : KAIST 전산학과 (석사)
2004년 : KAIST 전산학과 (박사)
2000년~2004년 : ㈜다음소프트 언어처리연구소

2004년~현 재 : 금오공과대학교 컴퓨터소프트웨어공학과 교수
※관심분야 : 자연언어처리, 텍스트데이터마이닝, 정보검색 등

김병만 (Byeong Man Kim)

1987년 : 서울대학교 컴퓨터공학과 (학사)
1989년 : KAIST 전산학과 (석사)
1992년 : KAIST 전산학과 (박사)

1992년~현 재 : 금오공과대학교 컴퓨터소프트웨어공학과 교수
※관심분야 : 인공지능, 정보검색, 컴퓨터보안, 정보보호 등