데이터 기반 러시아 대북 반응 조사

목차

[서론 3](#_Toc516665580)

[본문 4](#_Toc516665581)

[데이터 수집 4](#_Toc516665582)

[언론 반응 수집 4](#_Toc516665583)

[대중 반응 수집 5](#_Toc516665584)

[데이터 처리 및 모델링 5](#_Toc516665585)

[태그 추출 5](#_Toc516665586)

[감성 사전 6](#_Toc516665587)

[Vectorize Text 7](#_Toc516665588)

[BRNNs 9](#_Toc516665589)

[FastText 10](#_Toc516665590)

[Multinomial Naïve Bayes Classifier 11](#_Toc516665591)

[Logistic Classifier 12](#_Toc516665592)

[최종 모델 선정 12](#_Toc516665593)

[데이터 분석 13](#_Toc516665594)

[GDELT Tone 13](#_Toc516665595)

[모델을 통한 예측 14](#_Toc516665596)

[결론 19](#_Toc516665597)

[부록 20](#_Toc516665598)

[참고 문헌 23](#_Toc516665599)

# 서론

2018년 김정은은 신년사를 통해 “미국 본토 전역이 우리의 핵 타격 사정권 안에 있으며 핵 단추가 내 사무실 책상 위에 항상 놓여 있다” 고 했다. 그러나 불과 2 개월이 지나 평창올림픽 이후, 북한은 유화 정책으로 전환하는 듯한 모습을 보이며, 남북정상회담 및 미북정상회담이 성사되었다.

정상회담 이후 북한에 대한 본격적인 외교적 행동이 취해질 때 다시 6자회담이 성사될 가능성이 크다고 예상된다. 미국, 일본, 중국의 경우 이들의 대북 태도는 노선이 비교적 또렷하고 국내에서 이들의 소식을 접하기도 쉽다. 그러나 러시아의 경우, 미국의 트럼프 당선 이후 이전과 달리 친미 노선을 택함으로써, 대북 태도가 명료하지 않으며 국내에서 러시아의 여론을 접하기가 어렵기에 데이터를 기반으로 러시아의 북핵 관련 여론을 조사하였다.

조사는 크게 러시아 언론의 반응, 러시아 대중의 반응 두가지로 나누어서 진행되었다. 러시아 언론의 경우 러시아의 대표적인 보수적인 언론사인 ‘AIF’와 진보적인 언론사인 ‘에코 모스크바’의 북한 관련 기사들을 수집하였고, GDELT[[1]](#footnote-1)프로젝트에 있는 러시아의 북한 관련 데이터를 기반으로 조사되었다. 러시아의 대중 반응은 소셜 네트워크 서비스인Twitter와 각 언론사의 기사에 있는 코멘트를 수집하여 분석이 이루어졌다. 전자는 후자에 비해 객관적이고 공적인 매체이기에 현재 러시아 정부의 대북 태도를 조사하기 용이하다고 판단하였다. 반면 후자의 경우 대표성은 없으나 많은 데이터가 확보될 경우 대표성을 재고 시킬 수 있으며 실제 정치에 영향을 행사하는 것은 후자이기에 러시아의 장기적인 대북 태도를 유추하는 데에 도움이 될 것이라 예상하였다.

본 논문에 쓰인 모든 자료와 코드는 Github에 공개하였다. 향후 북한에 대한 새로운 여론 조사를 위하여 자세한 자료 수집과 조사에 관한 코드가 필요하다면, Github repository HYU-IECapstoneDesign-RussianTextMining/NorthKoreaReactionInRussia[[2]](#footnote-2)를 방문하면 얻을 수 있다.

# 본문

## 데이터 수집

### 언론 반응 수집

북한을 뜻하는 러시아어 “СеверНая Корея”와 미사일을 뜻하는 러시아어 “ракета”를 키워드로 러시아의 보수적 언론사인 AIF 진보적 언론사인 에코 모스크바의 기사를 검색하여 기사 본문과 이에 해당하는 코멘트를 수집하였다. 각 기사의 본문과 코멘트는 HTML을 파싱 하여 얻었으며 이는 python 라이브러리인 “beautiful soup”을 통하여 이루어 졌다.

러시아의 동향을 알기 위하여 GDELT 프로젝트의 데이터를 이용하였다. GDELT는 전세계의 기사를 수집 및 정리하는 프로젝트로 크게 EVENT 데이터와 Global Knowledge Graph로 이루어져 있다.

GDELT 2.0 EVENT 데이터베이스는 현재 2015 년 2 월부터 현재까지 15 분마다 업데이트된다. GDELT Translingual을 활용한 번역 기술로 65 개 언어의 기사를 모니터링 후 영어로 번역한다. 이후 이를 재가공하여 각 기사의 주제, 주체, 객체,지역, 감정 등을 제공한다.

GDELT 2.0 EVENT데이터베이스를 보완하는 GDELT 2.0 Global Knowledge Graph(이하 GKG) 는 단순한 기사 본문 텍스트를 넘어서 각 기사의 이미지, OCR 분석까지 제공하기에 EVENT 데이터 보다 더 풍부한 자료를 제공한다.

EVENT 데이터의 경우 Structured Data 이기에 구글의 Big Query Service를 통하여 쉽게 데이터를 얻을 수 있었다. 반면 GKG데이터는 데이터가 구조화되어 있지 않고 몇몇 필요한 데이터를 자체적으로 가공 후 Web API 형태로 JSON을 반환하는 형태였다. 또한 최근 3개월의 데이터만 검색할 수 있고 한 번 검색할 때 최대 250건의 데이터만 보여주는 한계가 있었다. 이러한 한계를 극복하기 위하여 R을 이용하여 주 단위로 요청을 하고 각 응답을 한 곳에 모아 data.frame을 반환하는 API를 재구성하였다.

### 대중 반응 수집

트위터 API와 함께 수작업으로 직접 대중의 반응을 수집하였다. 트위터 API의 경우, 한 번의 Request에 25개의 트윗만 반환하였고, 검색은 7일 이내로 제한되었다. 트윗을 반환할 때 중복된(리트윗된) 데이터가 상당히 많았기 때문에, API만으로는 충분한 양의 다양한 데이터를 확보할 수 없었다. 따라서, 이를 보조하기 위해 수작업으로 다음의 세 개의 단어를 직접 검색하여 데이터를 확보했다. 2018년 4월 11일부터 2018년 6월 6일 까지의 'Северная Корея(북한)', 'Пхеньян(평양)', 'Ким Чен Ына(김정은)' 단어를 포함한 트윗을 일자 별로 취합하여 분석에 사용하였다. 또한 기사를 수집할 때 함께 수집한 기사 댓글을 이용하여 언론사 별 소비자가 어떻게 다르게 반응하는지, 그리고 SNS 소비자와는 어떻게 다르게 반응하는지를 알아보았다.

## 데이터 처리 및 모델링

Twitter를 긍정, 부정으로 나눈 데이터, 그리고 전문가에게 도움을 받아 인터뷰를 긍정, 부정, 중립으로 분류한 데이터를 이용하였다.

전문가 데이터의 경우, 중립이 긍정/부정에 비해 6배 가량 많았기에 모델을 학습할 때에 중립으로 편향되고, 검증할 때에 또한 편향된 모델의 정확도가 높게 나왔다. 이 점을 해결하기 위해 전문가 데이터에 트위터의 긍정/부정 데이터를 혼합 데이터로 학습이 진행되었다.

### 태그 추출

러시아어 문장을 분석하기 위하여 데이터를 가공하는 과정에서 Part of Speech Tagging(이하 PoS)과 Lemmatisation(이하 Lemma)이 이용되었다. 본 논문에서는 R의 udpipe 라이브러리와 Universal Dependencies 에서 제공하는 UD Russian SynTagRus[[3]](#footnote-3) 모델이 이용되었으며, 해당 모델은 PoS에서 98%의 정확도, Lemma에서 90%의 정확도를 보인다고 알려져있다.(부록1)

PoS는 문장의 각 단어 별 품사를 결정하는 방법으로 제공되는 태그는 다음과 같다.

|  |  |  |
| --- | --- | --- |
| **Opened class words** | **Closed class words** | **Other** |
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CCONJ | X |
| NOUN | DET |  |
| PROPN | NUM |  |
| VERB | PART |  |
|  | PRON |  |
|  | SCONJ |  |

(자료 1)[[4]](#footnote-4)

Open class words에 해당하는 태그들은 문장에서 실질적인 의미를 가지는 품사들로 명사, 형용사, 동사와 같은 품사들이 해당되며, Closed class words는 조사, 정관사와 같은 실질적인 의미를 가지지 못하는 품사가 해당된다. 마지막으로 Other의 경우, 느낌표나 쉼표와 같은 품사에 해당하지 않는 것들이 해당된다.

Lemma는 각 단어의 표제어를 추출하는 방법으로, 단어의 어근을 추출하는 Stemming 기법과 유사한 방법이다. 그러나 Lemma는 Stemming과 달리 PoS까지 함께 고려한다. 예를 들어 ‘wound’가 동사로 사용되었을 경우 이는 ‘바람이 불다’의 과거형으로 wind로 처리되나 명사로 사용되었을 경우 상처라는 의미가 되기에 wound 그 자체로 처리되는 것이다.

### 감성 사전

감성 사전 방법은 감성 단어와 이에 해당하는 감성의 점수를 기록한 사전과 각 문장에 있는 단어들을 비교함으로써 각 문장의 점수를 계산하는 방식을 뜻한다. 러시아어의 경우 검증된 감정 사전이 존재하지 않아 부득이하게 UMass Lowell의 Text Machine Lab에서 이용된 러시아어 감성사전[[5]](#footnote-5)과 영어의 검증된 사전인 AFINN, Bing, Loghran, nrc를 러시아어로 Google Translate로 번역한 후 사용했다.

모든 데이터를 PoS, Lemma를 통해 처리하고, PoS 중 Opened class word만 추출하였다. 이후 사전 역시 Lemma 처리를 한 이후, 데이터와 사전을 비교하여 점수를 계산하였다.

|  |  |
| --- | --- |
| Twitter | Expert |
| 60 % | 60 % |

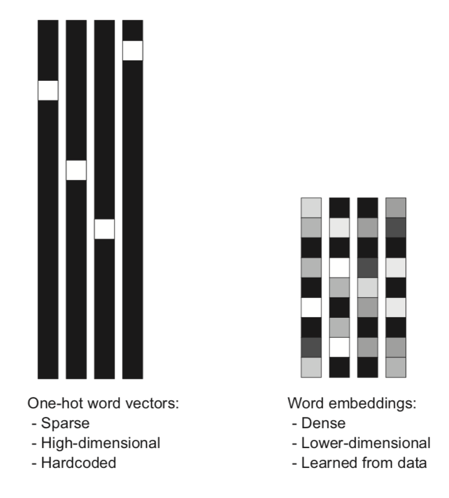
### Vectorize Text

러시아어 텍스트를 머신 러닝에 이용하기 위해 문장을 벡터화 할 필요가 있다.문장을 나누었을 때 하나의 뭉치를 토큰이라하며 이를 벡터화 하기 위한 기법으로 ‘one-hot encoding’, ‘token embedding’과 같은 기법들이 주로 사용된다.

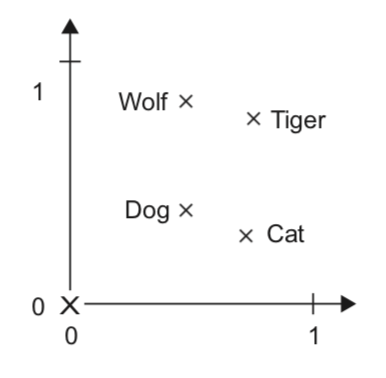
‘one-hot encoding’은 가장 많이 쓰이고, 가장 기본적인 벡터화 기법이다. 각 단어별로고유한 index를 부여한 후 이 index를 binary vector로 변환 후 단어가 있는 곳에만1을 부여하는 방법이다.(부록 1)

토큰을 벡터화 하는 또 다른 기법인 token embedding은 sparse하고(대부분이 0인) 고차원인 one-hot encoding 을 이용한 벡터 대신에 좀 더 조밀하고 저차원의 벡터를 이룬다. 또한 binary 데이터가 아닌 floating-point 데이터로 이루어져있으며, 데이터 자체가 hard-coded되어 있지 않고 데이터를 기반으로 학습된 수치가 벡터에 들어간다.

token embedding 은 매우 많은 단어들을 다룰 때에도 1024 차원정도이나 one-hot encoding은 20,000 단어만을 포함하더라도 20,000 차원의 벡터가 된다. 즉, word embedding이 더 적은 차원에서 더 많은 자료를 담고 있는 것이다.(자료 2)

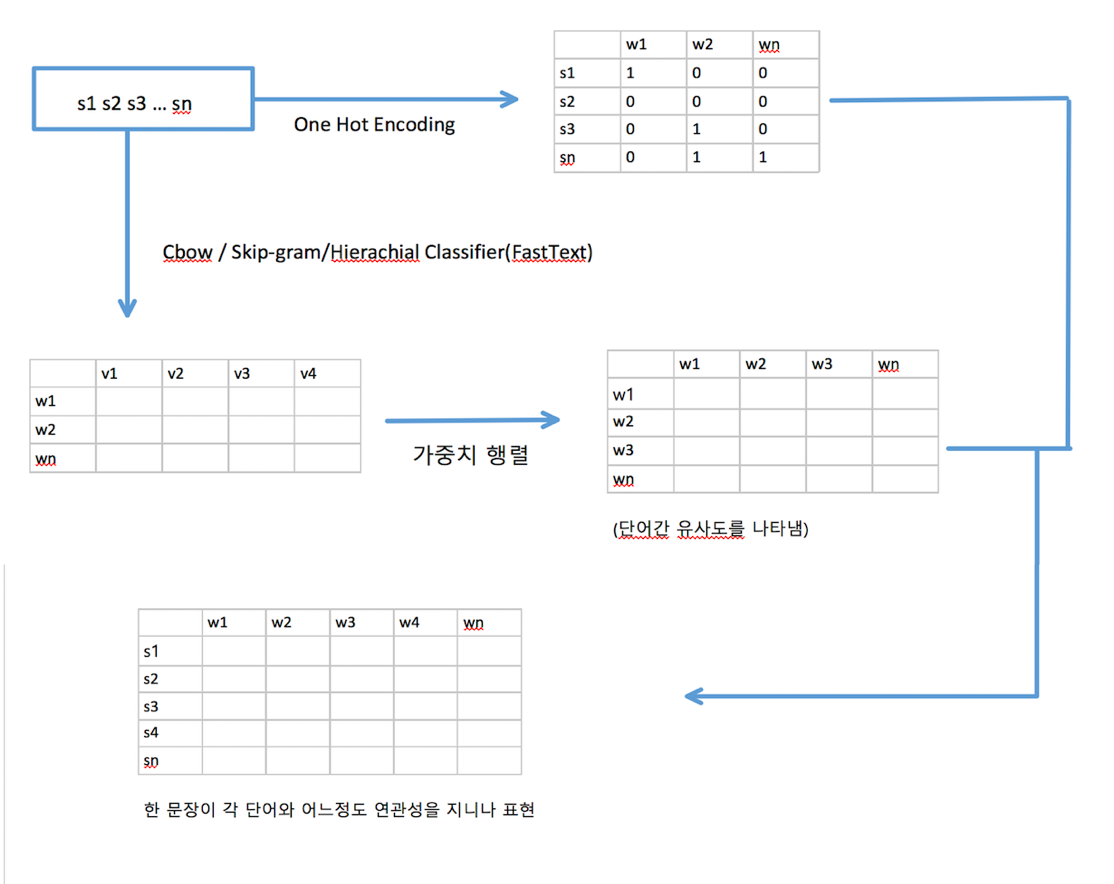
(자료2)[[6]](#footnote-6)

Token embedding은 인간의 언어를 기하학적 공간에 매핑한 것으로 거리가 가까운 단어는 서로 의미가 비슷한 단어를 방향은 공간에서 특정한 의미를 갖도록 벡터화 시키는 것이다. 예를 들어 (자료 3)과 같이 word embedding이 구성되어 있을 때, (wolf, tiger), (dog, cat)은 애완, 야생의 축으로 해석할 수 있고, (wolf, dog), (tiger, cat)은 개과, 고양이과로 해석할 수 있다.

(자료 3)[[7]](#footnote-7)

Word Embedding의 학습과정은 일반적인 인공 신경망의 학습과 비슷하다. 각 벡터에 이미 학습된 데이터 혹은 데이터를 초기화한 이후 이 값을 이용해 주변 문맥을 얼마나 정확하게 예측하는지 계산한다. 그리고 정확도가 좋지 못한 경우, 벡터의 값을 업데이트한다.

각 단어와 단어 사이의 연관성을 지닌 행렬과 one-hot-encoding의 결과를 곱함으로써, 한 문장이 한 단어와 얼마나 연관성을 지니는가를 알 수 있는데, 이 곱의 결과가 Input이 된다. (자료4)

(자료4)

### BRNNs

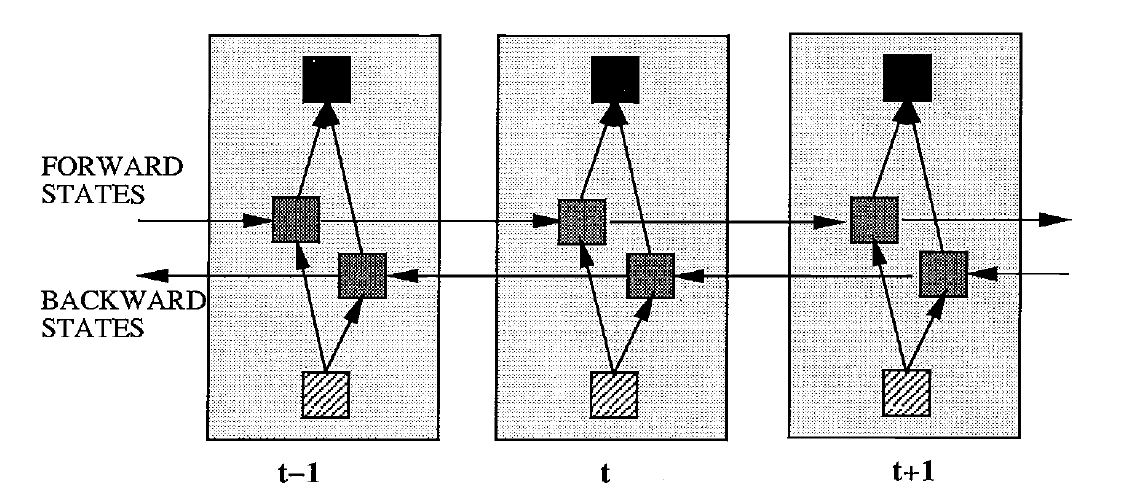
**RNN**

RNN(Recurrent Neural Network)은 일반적인 네트워크와 달리 현재 입력된 데이터와 과거에 입력 받았던 데이터를 동시에 고려한다.

(자료 5)[[8]](#footnote-8)

**Bidirectional RNN**

BRNN(Bidirectional Recurrent Neural Network)은 과거의 입력 데이터에 이후 정보까지 고려하는 네트워크이다.

(자료 6)[[9]](#footnote-9)

**Drop out**

드롭아웃 기법은 네트워크의 과적합현상을 막기위한 방법으로, 딥러닝 외의 알고리듬에서 사용되는 L1, L2Regularization 기법과 달리 훈련 시 랜덤하게 네트워크의 노드를 이용하지 않음으로써 과적합현상을 막는다.

**Activation**

Activation Function으로 sigmoid, ReLU, tanh가 이용되었다. tanh의 경우, RNN layer에서 ReLU의 경우, Fully Connected layer에서 sigmoid의 경우, output layer의 예측에 이용되었다.

본 논문에서는 500개의 단어, 200 단어의 문장으로 각 문장을 one-hot-encoding 처리 이후 embedded layer를 통하여 학습이후, bidirectional lstm layer에 fully connected layer를 연결하여 예측하는 네트워크를 구성하였다. 예측 결과는 validation set의 경우 65%로 나타났다. 38

(네트워크 구조)

### FastText

4번째 방법으로, Facebook에서 오픈소스로 제공하는 텍스트 분류 라이브러리 FastText를 이용하여 토큰화, word-embedding을 이용한 벡터화, 분류 예측을 하였다. FastText는 기존 word embedding 을 이용한 모델인 선형 분류기(Linear Classifier)가 아닌 계층 분류기(Hierachial Classifier)를 이용하는데 기존 방식 보다 빠르면서, 정확도는 딥러닝과 비슷한 수준을 낸다고 알려져 있다.(부록3)

선형분류기는 각 단어와 분류기준의 벡터를 비교하여 한 문장의 각 분류기준에 대한 확률을 계산하는 방식이다.’Cavs beat’, ‘Golden State’, ‘Serena Williams wins US Open’은 Sports에 대해서는 높은 확률을 보이겠으나, Travel에 대해서는 낮은 확률을 보일 것이므로 Sports로 분류할 수 있고, ‘Road trip’은 이와 반대이므로 Travel로 분류할 수 있다. 하지만 이 방법은 모든 각 단어를 모든 각 분류 기준에 대해 확률을 계산해야 되므로 계산량이 계층 분류기 보다 많을 수 밖에 없다.(자료 7)

(자료 7)[[10]](#footnote-10)

FastText에서 사용하는 계층 분류기는 분류를 Binary Tree로 구성함으로써 Sport 분류와 Championship분류는 비슷한 분류이고, 이 둘과 Travel은 동떨어진 분류이므로 우선 Travel이냐 아니냐를 분류하고, 그 이후에 Sport와 Championship을 분류하는 방식이다. 이와 같은 방식에서는 선형분류기와 달리 각 노드에 대해서만 확률을 계산하면 되기에 모델 훈련에 있어 시간이 절약된다.(자료 5)

(자료 8)[[11]](#footnote-11)

감성 사전 모델과의 비교를 위해 FastText에서도 역시 같은 데이터를 사용하였다.

72% / 55%

### Multinomial Naïve Bayes Classifier

다항분포 나이브 베이즈 분류기를 사용하였다. Java 기반 ML 어플리케이션 Weka를 사용하였으며 TF-IDF와 Snowball Stemmer를 이용하여 필요하지 않은 단어들을 제거 한 후 3-gram Word Token으로 예측하였다. 데이터의 경우, 앞서 사용한 Twitter 와 Expert를 혼합한 데이터를 이용하였고, 정확도는 Cross-validation의 경우 정확도 70%가 나타났다.

하지만 이 데이터의 경우 대부분의 긍부정이 Twitter 데이터에서 뽑은 것이기에 중립을 Expert 데이터로, 긍부정을 Twitter의 긍부정으로 편향되게 예측할 수 있다고 예상했기에, Expert데이터에서 긍정/부정/중립을 33개 씩 뽑아 예측을 하였는데 60%의 정확도로 나타났다.

### Logistic Classifier

나이브 베이즈 모델과 같이 해당 모델 역시 one-hot-encoding으로 처리한 문장을 input으로 예측한 모델이다. Python의 scikit-learn 패키지를 이용하였는데, 2-gram token으로 처리된 데이터를 이용한 결과 cross-validation 63 %, Expert 데이터에서 100개 추출한 데이터의 경우 90 %의 정확도가 나타났다.

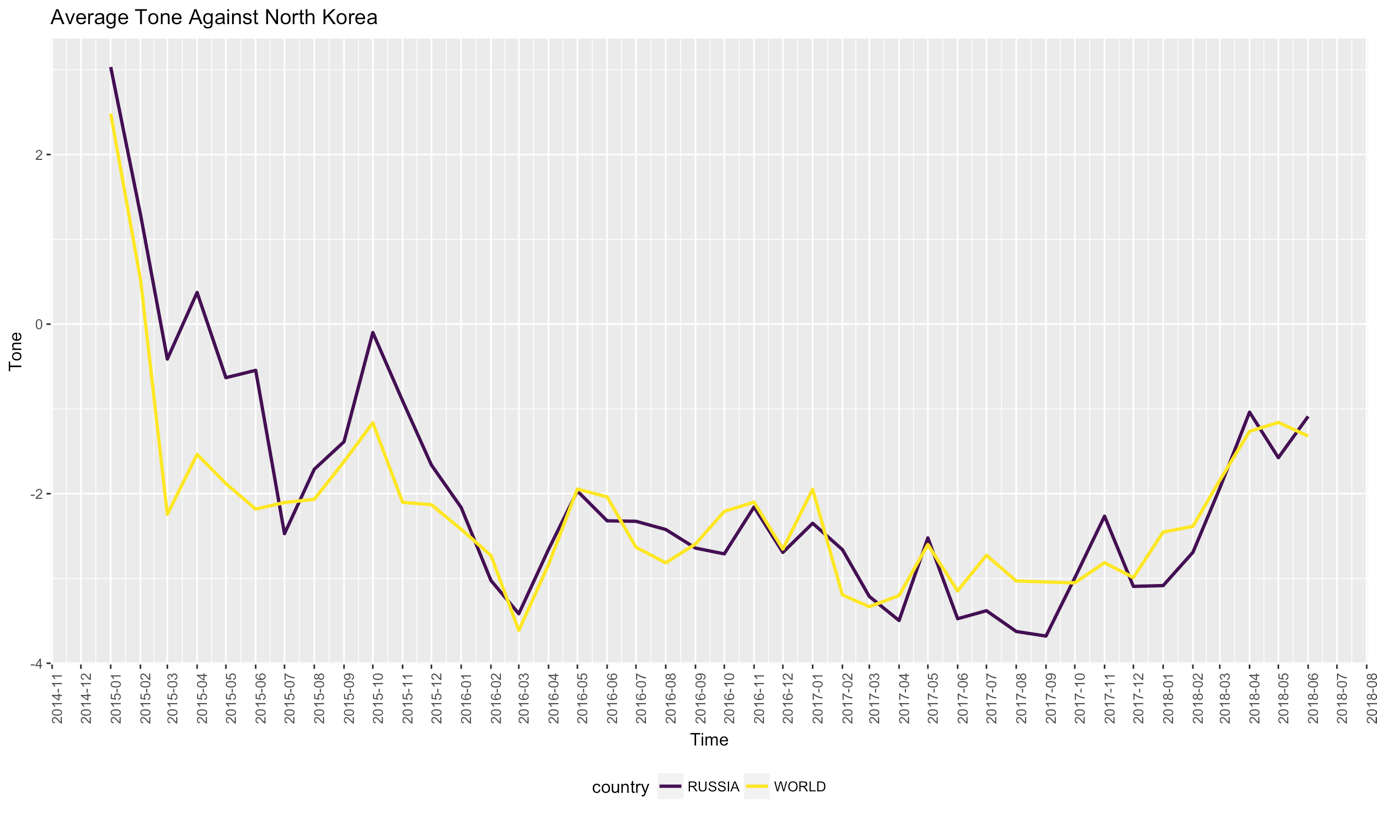
### 최종 모델 선정

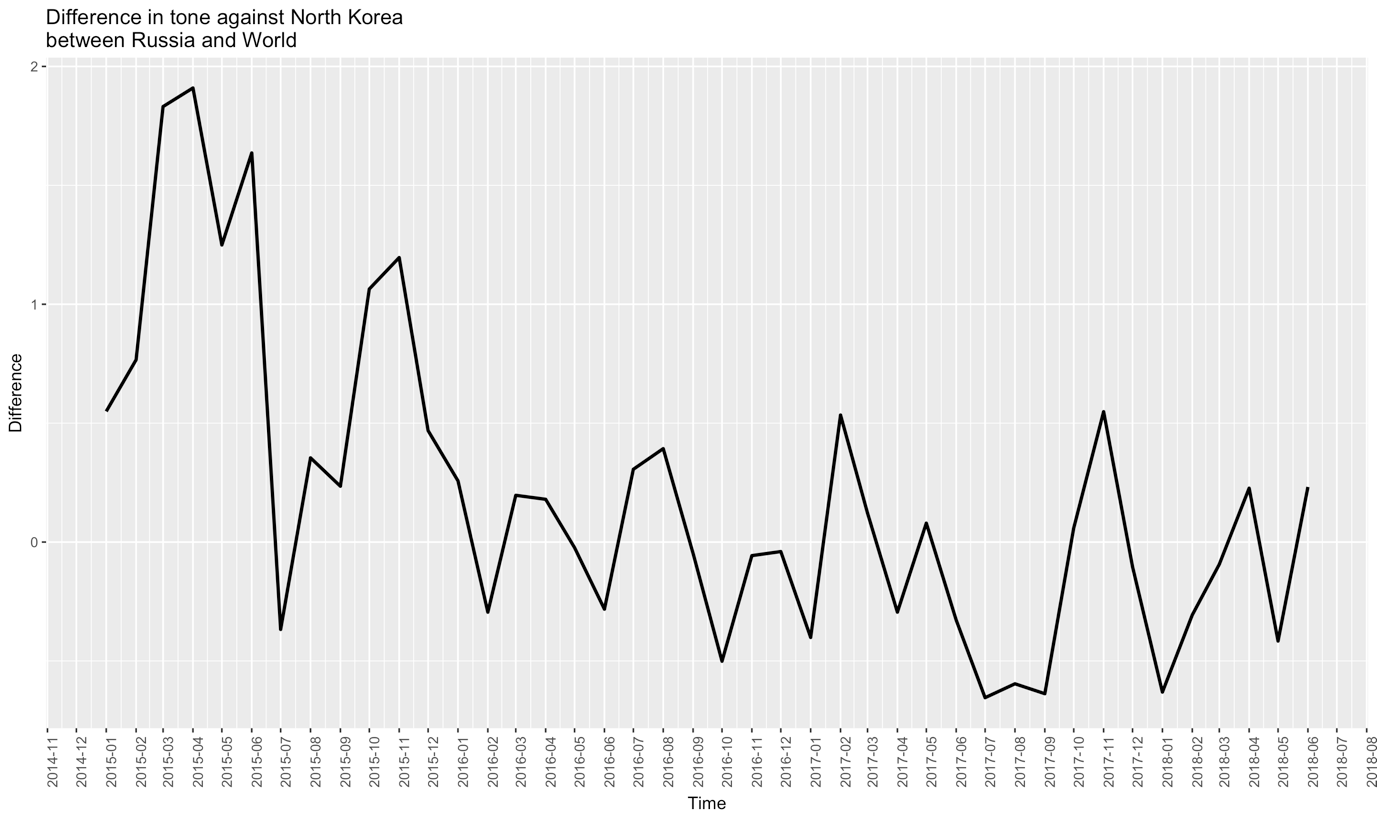
로지스틱 분류기가 가장 뛰어난 모델이었으나 좀 더 정밀한 판정을 위해, 전문가 데이터 600개를 추출하여, 각 문장을 번역기를 이용해 변형한 후 변형된 문장을 분류기를 통해 예측하였다. 정확도는 58%로 만족스러운 결과는 아니었으나, 긍정과 부정에 대한 정밀도(Precision)은 높은 수치가 나왔기에 해당 모델을 사용하기로 결정했다. (자료 9)

|  |  |  |
| --- | --- | --- |
| Threshold | Positive | Negative |
| 0.5 | 73% | 72% |
| 0.7 | 94% | 90% |
| 0.8 | 93% | 88% |

## 데이터 분석

### GDELT Tone

(자료7)

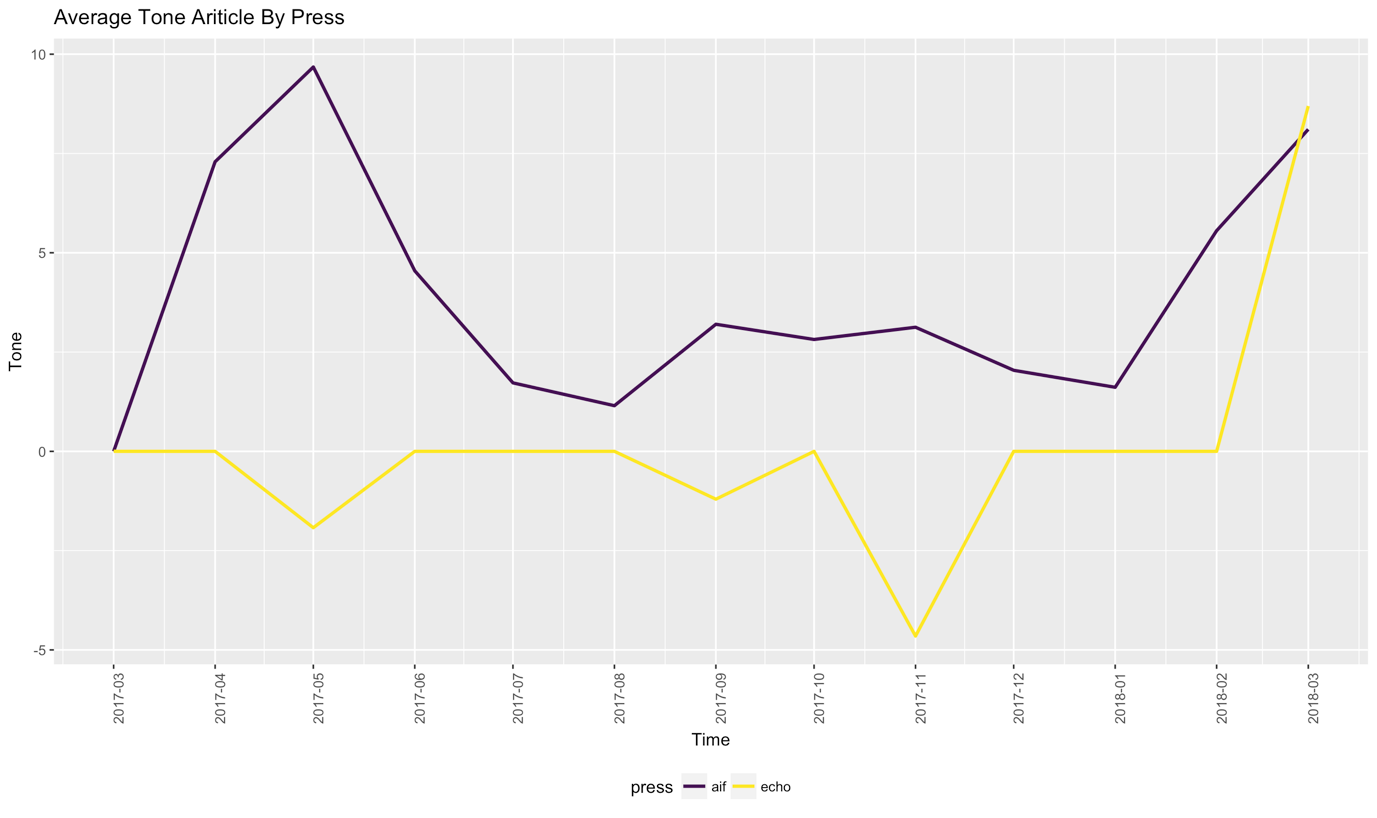
(자료 8)

GDELT의 EVENT데이터를 활용하여 북한 기사에 대한 Tone의 세계 평균과 러시아의 수치를 비교하였다. Tone의 경우 GDELT에서 각 기사 마다 자체적으로 계산한 값으로 -100 ~ 100의 점수를 제공하나, GDELT 측에서는 대게 -10 ~ 10의 점수를 지닌다고 설명하고 있다.[[12]](#footnote-12)

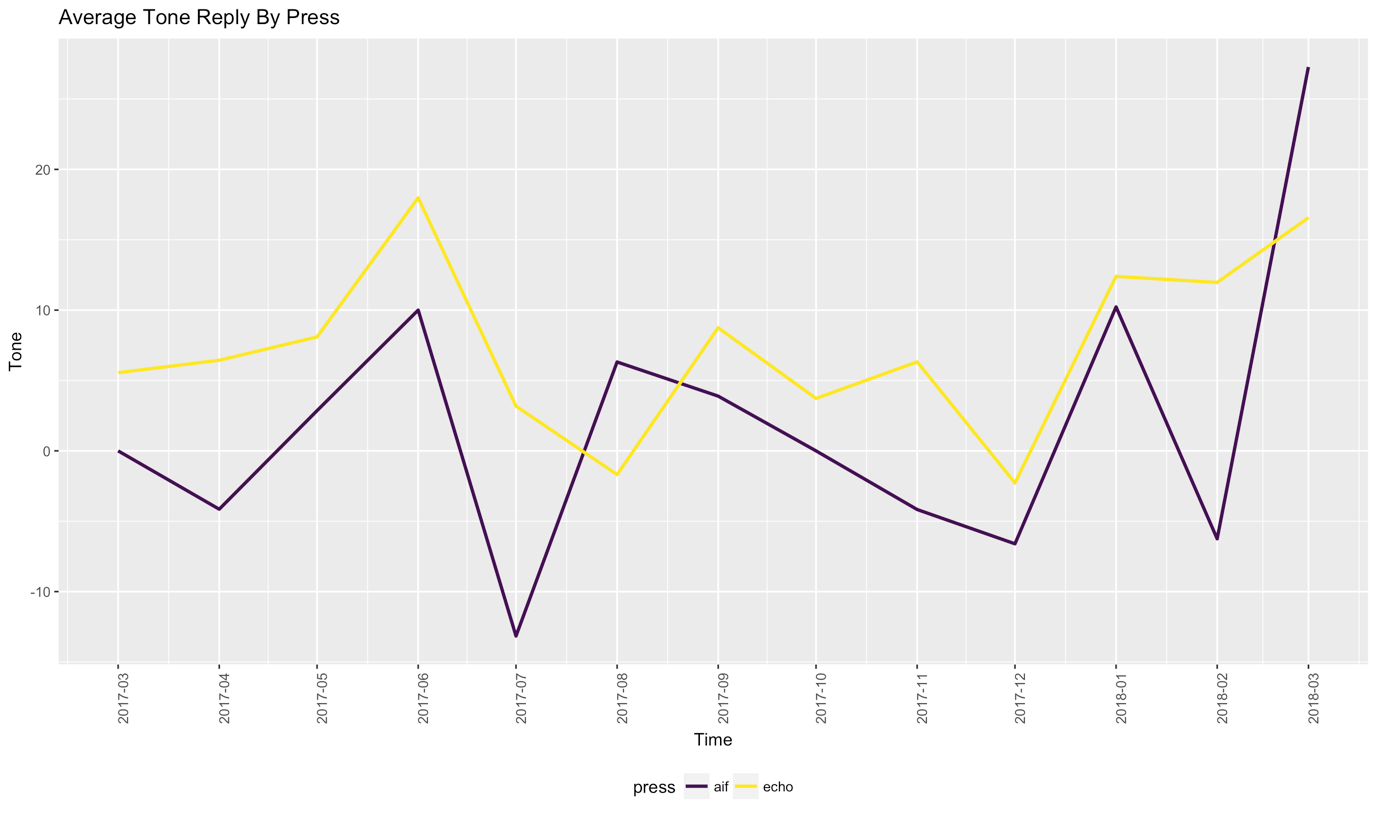
### 모델을 통한 예측

예측된 톤과 GDELT에서 제공하는 톤을 비교하였다. 전체적으로 비슷한 추이를 보이고 있으나, 좀 더 엄밀한 감정을 위하여 전문가에게 해당 자료를 보내어 검토하였다.

#### 대북 관련 기사를 통한 Tone 예측



#### 대북 관련 기사 댓글을 통한 Tone 예측



#### 대북 관련 트위터를 통한 Tone 예측

#### 대북 관련 기사 키워드

GKG에서 제공하는 기사의 제목을 이용하여 주간 대북 관련 기사의 중요 키워드를 뽑아내었다. 중요 키워드는 Miscrosoft Azure Text Analytics API에서 제공하는 핵심 문구 추출 API를 사용하였다. 핵심 문구의 번역은 Google Translate를 이용하였다.

|  |  |  |  |
| --- | --- | --- | --- |
| 날짜 |  | 핵심 문구 (러시아) | 핵심 문구 (영어) |
| 2018.1.1 | 2018.1.6 | Санкции | Sanctions |
| 2018.1.7 | 2018.1.13 | Президент Южной Кореи, северокорейский рыбак | President of South Korea, North Korean fisherman |
| 2018.1.14 | 2018.1.20 | Олимпиаду в Пхенчхан оркестр, Спортсмены из КНД и Южной Кореи | Olympics in Pyeongchang Orchestra, Athletes from CPV and South Korea, |
| 2018.1.21 | 2018.1.27 | Хоккеистки КНД, санкции США против КНД, КНД к участию в Олимпиаде | Hockey players of the CPV, US sanctions against the CPV, the CPV for participation in the Olympics |
| 2018.1.28 | 2018.2.3 | северокорейских рабочих в КНД в рамках санкций ООН, ракетах Северной Кореи, санкции ООН, Лыжники из Южной Кореи, леди Северной Кореи | North Korean workers in the CPV within the framework of UN sanctions, North Korea missiles, UN sanctions, Skiers from South Korea, North Korean ladies |
| 2018.2.4 | 2018.2.10 | Северокорейские артисты, Заход северокорейских судов в порты Южной Кореи, нарушением санкций | North Korean artists, Sunset of North Korean ships in the ports of South Korea, violation of sanctions |
| 2018.2.11 | 2018.2.17 | товарищ Ким, Президента Южной Кореи в Пхеньян, Иванка Трамп и Ким | Comrade Kim, the President of South Korea to Pyongyang, Ivanka Trump and Kim |
| 2018.2.18 | 2018.2.24 | младшая сестра Ким Чен Ына, мировой угрозой, сумму потраченных на визит делегации КНД на Олимпиаду денег | younger sister of Kim Jong-Yin, the world's threat, the amount of money spent on the visit of the CPV delegation to the Olympics |
| 2018.2.25 | 2018.3.3 | конструктивный диалог США и КНД, северокорейские болельщицы - клоны, сравнение с сестрой лидера КНД, секс - рабынями | constructive dialogue between the US and CPV, North Korean fans - clones, comparison with the leader's sister, sex slaves |
| 2018.3.4 | 2018.3.10 | X | X |
| 2018.3.11 | 2018.3.17 | США и Южная Корея, США и Северная Корея, Северную Корею к переговорам | USA and South Korea, the United States and North Korea, North Korea for talks |
| 2018.3.18 | 2018.3.24 | СШАи Южная Корея, столичном регионе, северокорейских беженцев, сотрудн-ичество в решении северокорейских проблем, переговоры о концертах южно-корейских артистов в КНД,биография супруги Ким Чен Ына, Украинский бизнес ресурс UBR, прямой секретный канал связи с Северной Кореей, Сон Хи и Чо Ён Пхиль, СеулиПхеньян, военныеучения, оссиейдлядостижения, Пхеньяне, Казах-стана, Сессия северокорейского парла-мента, северокорейские ракеты,Европы, Артисты из Южной Кореи, Известия, ua | USA and South Korea,the capital region,North Korean refugees,cooperation in solving North Korean problems,talks about concerts of South Korean artists in the CPV,biography of the wife of Kim Jong Yin, Ukrainian business resource UBR,a direct secret channel of communication with North Korea, Son Hee and Cho Yong Phil,Seoul and Pyong-yang,military training, |
| 2018.3.25 | 2018.3.31 | Поезд северокорейского лидера Ким Чен Ына, Пекин, северокорейских ракетах, северокорейских беженцев, Ким Чен Ына жирдяем в поезде в обход цензуре, Поезд с высокопоставленным официальным лицом из КНД, принципа развития отношений Китая и КНД, сотрудничество в решении северокорейских проблем, Севером и Югом Кореи, Музыканты из Южной Кореи, Отношение граждан США к северокорейской угрозе, Оружие, продажу ракетных технологий Северной Корее, столичном регионе | The train of the North Korean leader Kim Jong-un, Beijing, North Korean missiles, North Korean refugees, Kim Jong-Yin, wander the train bypassing the censorship, Train with a high-ranking official from the CPV, the principle of developing China-KNI relations, cooperation in solving North Korean problems, North and South Korea , Musicians from South Korea, The attitude of US citizens towards the North Korean threat, Weapons, the sale of missile technology to North Korea, the metropolitan region |
| 2018.4.1 | 2018.4.7 | Лидер КНД, санкции против КНД, концерт южнокорейских музыкантов в Пхеньяне, Ким Чен Ыну, Лидеру КНД, концерт южнокорейских поп - звезд, концерте южнокорейских артистов в Пхеньяне, концерт в присутствии Ким Чен Ына, концерт южнокорейских поп - музыкантов - Общество, концерт артистов из Южной Кореи, концерта в Пхеньяне, возможных переговоров США и КНД, восхищении концертом южнокорейских поп - музыкантов, южнокорейских девушек в юбках, ограничения для журналистов, США и Южная Корея, Лидер СК | Leader of the CPV, sanctions against the CPV, a concert of South Korean musicians in Pyongyang, Kim Jong-un, the leader of the CPV, a concert of South Korean pop stars, a concert of South Korean artists in Pyongyang, a concert in the presence of Kim Jong-no, a concert of South Korean pop musicians - Society, a concert of artists from South Korea, a concert in Pyongyang, possible negotiations between the US and CPV, admiration for a concert of South Korean pop musicians, South Korean girls in skirts, restrictions for journalists, the United States and South Korea, the leader of the UK |
| 2018.4.8 | 2018.4.14 | США и Северная Корея, тайные переговоры с США, Челябинской области, известная из жен северокорейских диктаторов, встречу Трампа с Ким Чен Ыном, Урал, готовности, Главы МИД Японии и Южной Кореи, взаимодействии, отказ от ядерного оружия | US and North Korea, secret negotiations with the US, Chelyabinsk region, known from the wives of North Korean dictators, a meeting between Trump and Kim Jong-no, Ural, readiness, the Foreign Minister of Japan and South Korea, cooperation, renunciation of nuclear weapons |
| 2018.4.15 | 2018.4.21 | Северная Корея, ядерные и ракетные испытания, проекту в КНД, Лидер КНД Ким Чен Ын, решение КНД, Представитель КНД, прекращение ядерных испытаний КНД, военные испытания, Южная Корея, ядерные испытания и запуски ракет | North Korea, nuclear and missile tests, the project in the CPV, the leader of the CPV Kim Jong-un, the decision of the CPV, the representative of the CPV, the termination of the nuclear tests of the CPR, military tests, South Korea, nuclear tests and missile launches |
| 2018.4.22 | 2018.4.28 | президента Южной Кореи, Ким Чен Ын, историческая встреча лидеров КНД и Южной Кореи, раунд переговоров между КНД и Южной Кореей, Северокорейские СМИ, ракетными запусками, содержании Пханмунчжомской декларации | President of South Korea, Kim Jong-un, a historic meeting of the leaders of the CPR and South Korea, a round of talks between the CPV and South Korea, the North Korean media, missile launches, the contents of the Panmunjom Declaration |
| 2018.4.29 | 2018.5.5 | Северная Корея, Южная Корея, визитом в КНД, Южной Кореей, Кореї, Мета Новини, Северную и Южную Корею, незначительный рост спроса на туры в Северную Корею | North Korea, South Korea, visit to the CPV, South Korea, Korea, Meta Novini, North and South Korea, insignificant growth in demand for tours to North Korea |
| 2018.5.6 | 2018.5.12 | Помпео, Северокорейские СМИ, электростанции в КНД, возобновлении экономических связей, северокорейский полигон Пхунгери, применения - Политика, Пхеньян - Владивосток - Сеул, авиасообщении с Южной Кореей, Южнокорейские студенты, северокорейском лидере, Южная Корея, Электроэнергетика, визитом в Северную Корею, Японию | Pompeo, North Korean media, power plants in CPV, renewal of economic ties, North Korean polygon Phungery, applications - Politics, Pyongyang - Vladivostok - Seoul, air communication with South Korea, South Korean students, North Korean leader, South Korea, Electricity, visit to North Korea, Japan |
| 2018.5.13 | 2018.5.19 | КНД в Южную Корею, Северной Кореи, танкер КНД, гражданами КНД, Южную Корею на лодочке, Военнослужащие Южной Кореи, Южную на лодочке, журналистов Южной Кореи на закрытие полигона, Береговая охрана Южной Кореи, Северную Корею, переговоры с Южной Кореей, диалог с Южной Кореей, Северная Корея, пути для возобновления диалога с КНД, демонтаж ядерного полигона Пхунгери, Южную Корею прекращением мирного диалога, саммит КНД и США под угрозой срыва, совместных учений с Южной Кореей, граждан КНД в южнокорейских водах, Северокорейский офицер на лодке | KND to South Korea, North Korea, CPV tanker, CPV citizens, South Korea on the boat, South Korean soldiers, South on the boat, South Korean journalists to close the landfill, South Korea Coast Guard, North Korea, talks with South Korea, dialogue with South Korea Korea, North Korea, ways for resuming dialogue with the CPR, dismantling the Phungeri nuclear test site, South Korea ending a peaceful dialogue, a summit of the CPV and the US under the threat of failure, joint exercises with South Korea, citizens of the CPV in South Korean waters, a North Korean officer on a boat |
| 2018.5.20 | 2018.5.26 | КНД на закрытие ядерного полигона, Закрытие северокорейского ядерного полигона, Южнокорейские журналисты, северокорейский Вонсан для освещения демонтажа ядерного полигона, рестораны в Камбодже | CPV for the closure of the nuclear test site, the closure of the North Korean nuclear test site, South Korean journalists, North Korean Wonsan for lighting the dismantlement of the nuclear test site, restaurants in Cambodia |
| 2018.5.27 | 2018.6.2 | Северную и Южную Корею URA, скором визите Лаврова в Северную Корею, Подготовка к встрече лидеров Северной Кореи и США, саммиту с Северной Кореей, встречей Трампа с Ким Чен Ыном | North and South Korea, the URA, the imminent visit of Lavrov to North Korea, the preparation for the meeting of the leaders of North Korea and the United States, the summit with North Korea, the meeting of Trump with Kim Jong-Eun |
| 2018.6.3 | 2018.6.9 | руководство КНД, связанных с КНД, Подготовка к саммиту КНД и США, размещение делегации КНД в Сингапуре, КНД с визитом - Политика, Северная Корея, лидера Северной Кореи, дорисованную улыбку Ким Чен Ына, Геополитическая сила Северной Кореи, специальную зону для встречи Трампа и Ким Чен Ына, Путина, Пропагандист Киселев, демонтаж на ключевом полигоне для ракетных испытаний | the leadership of the CPV related to the CPV, the preparation for the summit of the CPV and the US, the deployment of the CPV delegation in Singapore, the CPV on a visit - Politics, North Korea, the leader of North Korea, the drew smile of Kim Jong-no, the geopolitical force of North Korea, a special zone for the meeting of Trump and Kim Jong Yin, Putin, Propagandist Kiselev, dismantling at a key test range for missile testing |

(자료 9)

# 결론

# 부록

부록 1[[13]](#footnote-13)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Treebank** | **Mode** | **Words** | **Sents** | **UPOS** | **XPOS** | **Feats** | **AllTags** | **Lemma** | **UAS** | **LAS** |
| Russian | Raw text | 99.90% | 96.90% | 94.70% | 94.40% | 84.40% | 82.80% | 75.00% | 80.30% | 75.50% |
| Russian-SynTagRus | Raw text | 99.60% | 98.00% | 98.00% | - | 93.60% | - | 95.60% | 89.80% | 87.20% |

부록2[[14]](#footnote-14)

“The cat sat on the mat”

“The dog ate my homework”

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | The | cat | sat | on | the | ma  t | dog | ate | my | homework |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

, , 1

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 0 0 0 0 0 0

[2,] 0 0 0 0 0 0 0 0 0 0

, , 2

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 1 0 0 0 0 0 0 0 0 0

[2,] 1 0 0 0 0 0 0 0 0 0

, , 3

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 1 0 0 0 0 0 0 0 0

[2,] 0 0 0 0 0 0 0 0 0 0

, , 4

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 1 0 0 0 0 0 0 0

[2,] 0 0 0 0 0 0 0 0 0 0

, , 5

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 1 0 0 0 0 0 0

[2,] 0 0 0 0 0 0 0 0 0 0

, , 6

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 1 0 0 0 0 0

[2,] 0 0 0 0 0 0 0 0 0 0

, , 7

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 0 1 0 0 0 0

[2,] 0 0 0 0 0 0 0 0 0 0

, , 8

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 0 0 0 0 0 0

[2,] 0 1 0 0 0 0 0 0 0 0

, , 9

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 0 0 0 0 0 0

[2,] 0 0 1 0 0 0 0 0 0 0

, , 10

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 0 0 0 0 0 0

[2,] 0 0 0 1 0 0 0 0 0 0

, , 11

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 0 0 0 0 0 0 0 0 0 0

[2,] 0 0 0 0 1 0 0 0 0 0

부록3[[15]](#footnote-15)



# 참고 문헌

Charles University, Czech Republic. (n.d.). *UDPipe User's Manual.* Retrieved from Institute of Formal and Applied Linguistics: https://ufal.mff.cuni.cz/udpipe/users-manual

Facebook. (2018). *FastText*. Retrieved from fasttext.cc

François Chollet, J. A. (2018). *Deep Learning with R.* Manning Publications.

Romanov, A. (2017, 5 12). *A simple dictionary-based sentiment analysis system with Russian language support.* Retrieved from Github: https://github.com/text-machine-lab/sentimental/blob/master/sentimental/word\_list/russian.csv

Scharfstein, M. (2017, 9 20). *FastText Tutorial - How to Classify Text with FastText .* Retrieved from Youtube: https://www.youtube.com/watch?v=4l\_At3oalzk

The GDELT Project. (n.d.). *The GDELT Project.* Retrieved from https://www.gdeltproject.org

Universal Dependencies. (2017, 3). *Universal Dependencies.* Retrieved from http://universaldependencies.org

1. https://www.gdeltproject.org/ (The GDELT Project) [↑](#footnote-ref-1)
2. https://github.com/HYU-IECapstoneDesign-RussianTextMining/NorthKoreaReactionInRussia [↑](#footnote-ref-2)
3. http://universaldependencies.org/treebanks/ru\_syntagrus (Universal Dependencies, 2017) [↑](#footnote-ref-3)
4. http://universaldependencies.org/u/pos/ (Universal Dependencies, 2017) [↑](#footnote-ref-4)
5. https://github.com/text-machine-lab/sentimental/blob/master/sentimental/word\_list/russian.csv (Romanov, 2017) [↑](#footnote-ref-5)
6. François Chollet, J. A., *Deep Learning with R*, 2018, p169 [↑](#footnote-ref-6)
7. François Chollet, J. A., *Deep Learning with R*, 2018, p170 [↑](#footnote-ref-7)
8. https://web.stanford.edu/group/pdplab/pdphandbook/handbookch8.html [↑](#footnote-ref-8)
9. http://solarisailab.com/archives/1515 [↑](#footnote-ref-9)
10. https://www.youtube.com/watch?v=4l\_At3oalzk (Scharfstein, 2017) [↑](#footnote-ref-10)
11. https://www.youtube.com/watch?v=4l\_At3oalzk (Scharfstein, 2017) [↑](#footnote-ref-11)
12. http://data.gdeltproject.org/documentation/GDELT-Event\_Codebook-V2.0.pdf, p5 [↑](#footnote-ref-12)
13. https://ufal.mff.cuni.cz/udpipe/users-manual (Universal Dependencies, 2017) [↑](#footnote-ref-13)
14. François Chollet, J. A., *Deep Learning with R*, 2018, p167 [↑](#footnote-ref-14)
15. https://fasttext.cc/blog/2016/08/18/blogpost.html [↑](#footnote-ref-15)