데이터 기반 러시아 대북 반응 조사

목차

[서론 2](#_Toc510304155)

[본문 3](#_Toc510304156)

[데이터 수집 3](#_Toc510304157)

[언론 반응 수집 3](#_Toc510304158)

[대중 반응 수집 4](#_Toc510304159)

[데이터 처리 및 모델링 5](#_Toc510304160)

[감성 사전 5](#_Toc510304161)

[Miscrosoft Azure Text Analytics API 6](#_Toc510304162)

[Deep Learning 6](#_Toc510304163)

[FastText 8](#_Toc510304164)

[GDELT GCAM 9](#_Toc510304165)

[GDELT GCAM Azure API, FastText 모델 비교 10](#_Toc510304166)

[구어체/문어체 분류기 11](#_Toc510304167)

[최종 모델 11](#_Toc510304168)

[데이터 분석 11](#_Toc510304169)

[언론 반응 11](#_Toc510304170)

[대중 반응 15](#_Toc510304171)

[부록 16](#_Toc510304172)

[참고 문헌 19](#_Toc510304173)

# 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 거의 성사된 것으로 보인다.

정상회담이 성사된 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 ‘이즈베스티야’와 진보적인 언론사인 ‘노비야 가제타’의 북한 관련 기사들을 수집하였고, GDELT[[1]](#footnote-1)프로젝트에 있는 러시아의 북한 관련 데이터를 기반으로 조사되었다. 러시아의 대중 반응은 소셜 네트워크 서비스인 vk, Facebook, Twitter를 통하여 데이터와 각 언론사의 기사에 있는 코멘트를 수집하였다. 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 jjwkdl/ 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 이즈베스티야와 진보적 언론사인 노바야 가제타의 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. 북한 미사일 관련 데이터는 핵안보 관련 단체인 NTI에서 제공하는 자료[[3]](#footnote-3)를 수집하였다.

러시아의 대표적 언론뿐만 아니라 러시아의 전체적인 언론 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 구글의Big 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

## 데이터 처리 및 모델링

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

Twitter 데이터의 경우, 텍스트 내용의 긍/부정 보다는 화자의 감정을 통해 분류된 데이터였고, 나머지 데이터의 경우, 텍스트 내용을 기반으로 감정을 분류한 데이터이기에 여러 데이터를 합쳐서 학습을 훈련시키는 것이 부적절 하다고 판단하였다. 그렇기에, 각 데이터를 통해 학습 한 후 각 데이터를 검증하는 방식으로 진행하였다.

### 감성 사전

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[[4]](#footnote-4) 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)[[5]](#footnote-5)

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 방법은 감성 단어와 이에 해당하는 감성의 점수를 기록한 사전과 각 문장에 있는 단어들을 비교함으로써 각 문장의 점수를 계산하는 방식을 뜻한다. 영어의 경우, AFINN과 같은 검증이 된 감정 사전이 존재하나, 러시아어의 경우 검증된 감정 사전이 존재하지 않아 부득이하게 UMass Lowell의 Text Machine Lab에서 이용된 러시아어 감성사전[[6]](#footnote-6)을 이용하였다.

감정 라벨이 있는 문장은 A에서 얻었으며 이를 모두 PoS, Lemma를 통해 처리하고, PoS 중 Opened class word만 추출하였다. 이후 사전 역시 Lemma 처리를 한 이후, 데이터와 사전을 비교하여 점수를 계산하였다.

|  |  |  |
| --- | --- | --- |
| Kaggle | Twitter | Expert |
| 43 % | 60 % | 55 % |

### Miscrosoft Azure Text Analytics API

Microsoft Azure는 cognitive 서비스 중 텍스트 분석 API를 제공하고 있다. 유료 API이기에 자세한 기술 사양은 알 수 없었으나 Google의 Cloud NLP API와 IBM의 Watson API와 달리 러시아어를 지원하기에 이용해 보았다. 이 API역시 Web API로 JSON파일을 보내면 결과를 JSON으로 받는 형식이었기에, 다시 R에 맞게 data.frame을 주고 받을 수 있도록 API를 재설계 하였다.

|  |  |  |
| --- | --- | --- |
| Kaggle | Twitter | Expert |
| 40 % | 75.12 % | 67.46 % |

### Deep Learning

**Word Embedding**

러시아어 텍스트를 머신 러닝에 이용하기 위해 문장을 벡터화 할 필요가 있다.문장을 나

누었을 때 하나의 뭉치를 토큰이라하며 이를 벡터화 하기 위한 기법으로 ‘one-hot encoding’, ‘token embedding’과 같은 기법들이 주로 사용된다.

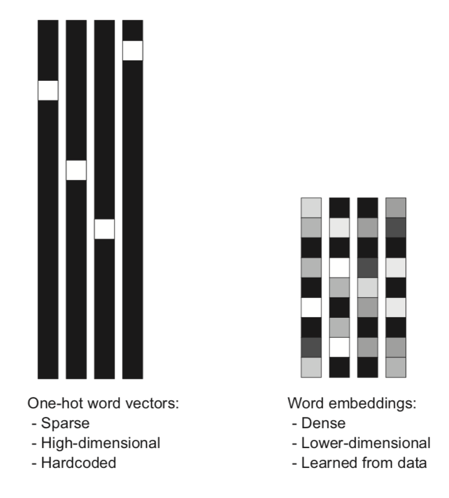
‘one-hot encoding’은 가장 많이 쓰이고, 가장 기본적인 벡터화 기법이다. 각 단어 별로

고유한 index를 부여한 후 이 index를 binary vector로 변환 후 단어가 있는 곳에만1을 부여

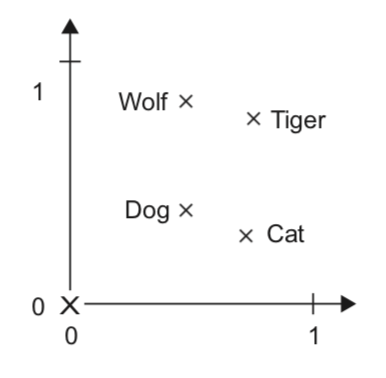
하는 방법이다.(부록 1)

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 embe-dding이 더 적은 차원에서 더 많은 자료를 담고 있는 것이다.(자료 2)

(자료2)[[7]](#footnote-7)

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)[[8]](#footnote-8)

**RNN-GRU**

**Conv 1d (filter kernel padding stride)**

**Max pooling**

**Drop out (overfit)**

**Bidirectional**

**Activation (logistic, relu, tanh) 구어체72%**

**좌모델 상데이터**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Twitter (Binary) | Kaggle | Expert |
| Twitter (Binary) | 62.3 % | 31.1 % | 43.6 % |
| Kaggle | 51.1 % | 55.8 % | 39.3 % |
| Expert | 49.7 % | 48.5 % | 70.5 % |

**From Twitter To Kaggle, Expert -> using threshold 1/3 2/3 3/3**

**From Kaggle, Expert To Twitter -> train with only negative and positive**

### FastText

4번째 방법으로, Facebook에서 오픈소스로 제공하는 텍스트 분류 라이브러리 FastText

를 이용하여 토큰화, word-embedding을 이용한 벡터화, 분류 예측을 하였다. FastText는 기존 word embedding 을 이용한 모델인 선형 분류기(Linear Classifier)가 아닌 계층 분류기(Hiera-chial 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 4)

(자료 4)[[9]](#footnote-9)

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 5)[[10]](#footnote-10)

감성 사전 모델과의 비교를 위해 FastText에서도 역시 같은 데이터를 사용하였다.

좌 모델

|  |  |  |  |
| --- | --- | --- | --- |
|  | Twitter (Binary) | Kaggle | Expert |
| Twitter (Binary) | 82 % | X | X |
| Kaggle | X | 66 % | 49 % |
| Expert | X | 46 % | 73 % |

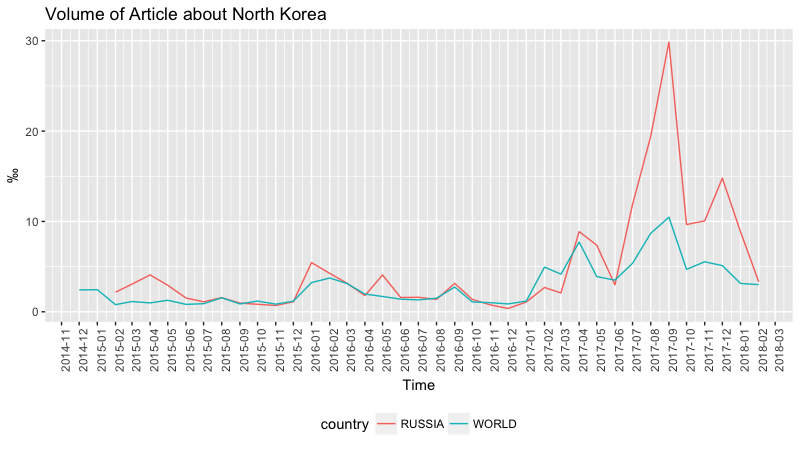
### GDELT GCAM

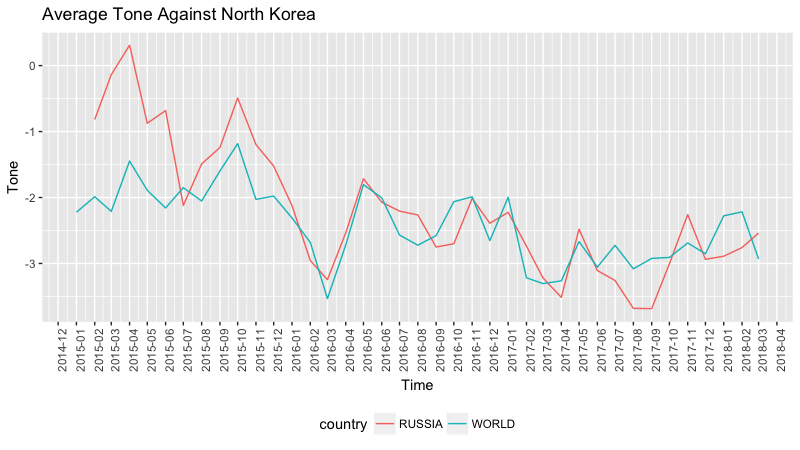
### 최종 모델

## 데이터 분석

### 언론 반응

#### 대북 관련 기사의 비율과 경향

(자료6)

(자료7)

(자료 8)

GDELT의 EVENT데이터를 활용하여 북한 기사에 대한 Volume과 Tone의 세계 평균과 러시아의 수치를 비교하였다. 여기서 Volume은 일정 기간의 전체 기사 수 중 북한 관련 기사의 비중을 나타내는 것으로 퍼밀(‰) 단위로 계측하였다. Tone의 경우 GDELT에서 각 기사 마다 자체적으로 계산한 값으로 -100 ~ 100의 점수를 내나, GDELT 측에서는 대게 -10 ~ 10의 점수를 지닌다고 설명하고 있다.[[11]](#footnote-11)

**(제출 전 업데이트)**

**(새로운 구글 클라우드 계정 필요)**

2016 09 북 SLBM, 한러 정상회담

2017 07 북러이 패키지 제재 미 의회 통과

2017 09 북 6차 핵실험

#### 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 |  |  |
| 2018.4.1 | 2018.4.7 |  |  |
| 2018.4.8 | 2018.4.14 |  |  |
| 2018.4.15 | 2018.4.21 |  |  |
| 2018.4.22 | 2018.4.28 |  |  |
| 2018.4.29 | 2018.5.5 |  |  |
| 2018.5.6 | 2018.5.12 |  |  |
| 2018.5.13 | 2018.5.19 |  |  |
| 2018.5.20 | 2018.5.26 |  |  |
| 2018.5.27 | 2018.6.2 |  |  |
| 2018.6.3 | 2018.6.9 |  |  |

(자료 9)

Title, Mood(GDELT, FT, Azure) (X) 동시 추출 안됨

#### 언론사별 반응

### 대중 반응

#### 언론사별 댓글

#### Twitter

# 부록

부록 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| url | title | gdelt | azure | Fast  text |
| http://gursesintour.com/aktualnye-novosti/flash-player-raskritikovali-za-uyazvimosti-dlya-hakerov/1156282/ | Flash Player раскритиковали за уязвимости для хакеров | -18 | 0.53 | 1  0.84 |
| https://glavnoe.ua/news/n334927-ja-mstju-i-mstja-moja-strashna-rossijskie-hakery-atakovali-olimpiadu | Я мстю , и мстя моя страшна : российские хакеры атаковали Олимпиаду | -15 | 0.51 | 0  0.93 |
| http://inforos.ru/?module=news&action=view&id=63496 | Косачев : санкции США против КНД провоцируют срыв урегулирования ситуации на полуострове / Новости / Информационное агентство Инфорос | -14 | 0.47 | 0  0.74 |
| http://argumenti.ru/crime/2018/01/561759 | В хостеле Москвы произошла драка со стрельбой , пострадали северокорейцы | -13 | 0.48 | 0  0.60 |
| http://internetua.com/haker-iz-kndr-atakovali-krupneishuua-energeticseskuua-kompaniua-izrailya | Хакеры из КНД атаковали крупнейшую энергетическую компанию Израиля | -12 | 0.57 | 1  0.93 |
| http://world.kbs.co.kr/russian/./news/news\_In\_detail.htm?No=51898&id=In | Северокорейские хакеры стали мировой угрозой | -11 | 0.52 | 0  0.81 |
| https://iz.ru/712483/2018-02-23/dva-cheloveka-zaderzhany-za-ataku-na-assotciatciiu-severokoreiskikh-grazhdan-v-tokio | Два человека задержаны за атаку на Ассоциацию северокорейских граждан в Токио | Новости | Известия | -10 | 0.52 | 0  0.95 |
| https://regnum.ru/news/polit/2368965.html | Трамп : оссия нарушает режим международных санкций против Северной Кореи | -9 | 0.47 | 0, 0.95 |
| http://internetua.com/severokoreiskaya-hakerskaya-organizaciya-reaper-stanovitsya-globalnoi-ugrozoi | Северокорейская хакерская организация Reaper становится глобальной угрозой | -8 | 0.54 | 0, 0.99 |
| https://telegraf.com.ua/mir/usa/3872174-ssha-rasshirili-sanktsii-protiv-severnoy-korei.html | США расширили санкции против Северной Кореи - Новости США - В черный список попали 9 компаний , 16 физических лиц и 6 северокорейских судов | -7 | 0.46 | 1, 0.94 |
| https://telegraf.com.ua/mir/aziya/3930457-delegatsiya-kndr-so-skandalom-pribyila-na-zakryitie-olimpiadyi.html | Делегация КНД со скандалом прибыла на закрытие Олимпиады - Новости Азии - яд депутатов от южнокорейских оппозиционных партий устроили сидячую забастовку возле границы с Северной Кореей | -6 | 0.56 | 1, 0.97 |
| https://echo.msk.ru/news/2125688-echo.html | Высокопоставленная северокорейская делегация возможно примет участие в церемониях открытия и закрытия Олимпиады Новости Эхо Москвы , 09 . 01 . 2018 | 6 | 0.58 | 1, 0.99 |
| https://lenta.ru/news/2018/02/22/northkorea/ | Северокорейской болельщице досталось за аплодисменты американцам : Зимние виды : Спорт : Lenta . ru | 7 | 0.57 | 0, 0.99 |
| https://nv.ua/world/countries/na-olimpiadu-v-juzhnuju-koreju-pribyli-bolelshchitsy-iz-kndr-2450361.html | На Олимпиаду в Южную Корею прибыли болельщицы из КНД / Новое Время | 8 | 0.54 | 0, 0.61 |
| http://www.kazpravda.kz/articles/view/200-idealnih-bolelshchits-otpravila-na-olimpiadu-severnaya-koreya/ | 200 идеальных болельщиц отправила на Олимпиаду Северная Корея - Новости Казахстана | 9 | 0.47 | 1, 0.94 |
| https://russian.rt.com/world/news/497348-si-czinpin-vizit-kndr | СМИ : Си Цзиньпин принял приглашение Ким Чен Ына посетить КНД Т на русском | 11 | 0.58 | 1, 0.58 |
| https://telegraf.com.ua/mir/drugie/3888068-iordaniya-razorvala-diplomaticheskie-otnosheniya-s-kndr.html | Иордания разорвала дипломатические отношения с КНД - Другие новости мира - ешение о разрыве дипотношений с КНД утверждено указом короля Абдаллы II | 13 | 0.53 | 0, 0.74 |

부록 2[[12]](#footnote-12)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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% |

부록3[[13]](#footnote-13)

“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

부록4[[14]](#footnote-14)



# 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

CNS. (n.d.). *NTI.* Retrieved from The CNS North Korea Missile Test Database: http://www.nti.org/analysis/articles/cns-north-korea-missile-test-database

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

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

Gentry, J. (2015, 7 29). *twitteR: R Based Twitter Client.* Retrieved from CRAN: https://cran.r-project.org/web/packages/twitteR

Hester, J. (2017, 10 29). *glue: Interpreted String Literals.* Retrieved from CRAN: https://cran.r-project.org/web/packages/glue

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

Wickham, H. (2017, 11 4). *tidyverse: Easily Install and Load the 'Tidyverse'.* Retrieved from CRAN: https://cran.r-project.org/web/packages/tidyverse

Wijffels, J. (2018, 3 12). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit.* Retrieved from CRAN: https://cran.r-project.org/web/packages/udpipe

1. https://www.gdeltproject.org/ (The GDELT Project) [↑](#footnote-ref-1)
2. https://github.com/jjwkdl/ NorthKoreaReactionInRussia [↑](#footnote-ref-2)
3. https://www.nti.org/documents/2137/north\_korea\_missile\_test\_database.xlsx (CNS) [↑](#footnote-ref-3)
4. http://universaldependencies.org/treebanks/ru\_syntagrus (Universal Dependencies, 2017) [↑](#footnote-ref-4)
5. http://universaldependencies.org/u/pos/ (Universal Dependencies, 2017) [↑](#footnote-ref-5)
6. https://github.com/text-machine-lab/sentimental/blob/master/sentimental/word\_list/russian.csv (Romanov, 2017) [↑](#footnote-ref-6)
7. François Chollet, J. A., *Deep Learning with R*, 2018, p169 [↑](#footnote-ref-7)
8. François Chollet, J. A., *Deep Learning with R*, 2018, p170 [↑](#footnote-ref-8)
9. https://www.youtube.com/watch?v=4l\_At3oalzk (Scharfstein, 2017) [↑](#footnote-ref-9)
10. https://www.youtube.com/watch?v=4l\_At3oalzk (Scharfstein, 2017) [↑](#footnote-ref-10)
11. http://data.gdeltproject.org/documentation/GDELT-Event\_Codebook-V2.0.pdf, p5 [↑](#footnote-ref-11)
12. https://ufal.mff.cuni.cz/udpipe/users-manual (Universal Dependencies, 2017) [↑](#footnote-ref-12)
13. François Chollet, J. A., *Deep Learning with R*, 2018, p167 [↑](#footnote-ref-13)
14. https://fasttext.cc/blog/2016/08/18/blog-post.html [↑](#footnote-ref-14)