NAZARBAYEV UNIVERSITY

Sentiment analysis & visualization of data from social networks using Machine learning algorithms

Student: Aru Omarali

Supervisor: Askar Boranbayev

Co-Supervisor: Mark Sterling

Date: 29.04.2021

NAZARBAYEV UNIVERSITY

Content

- Background
- Motivation
- Objectives
- Modern approaches
- Literature review
- Methodology
- Experimental work
- Results
- Data visualization
- Discussion
- Application
- Conclusion
- Future work

NAZARBAYEV UNIVERSITY

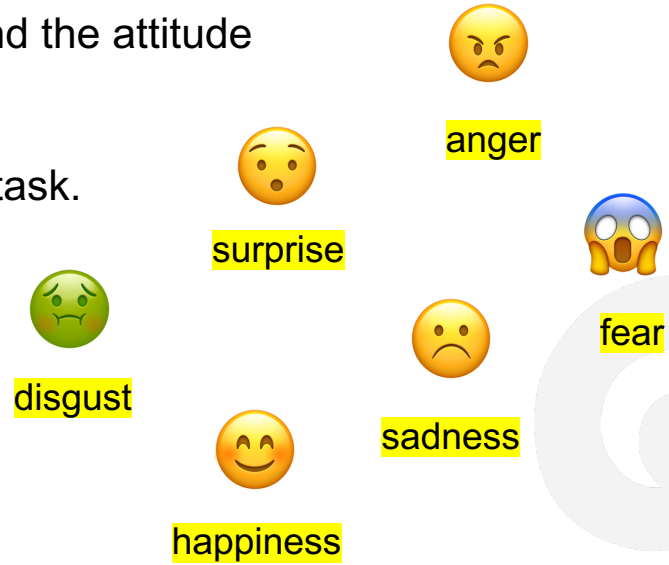
Sentiment analysis

What is it?

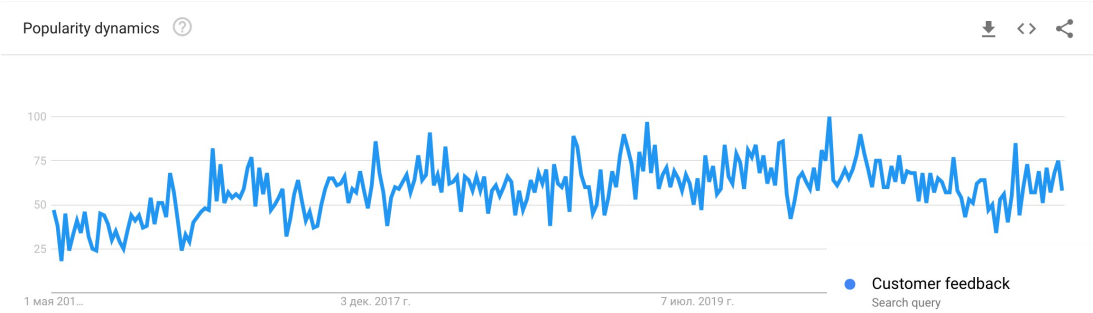
Process of defining emotions from text to understand the attitude towards some concept.

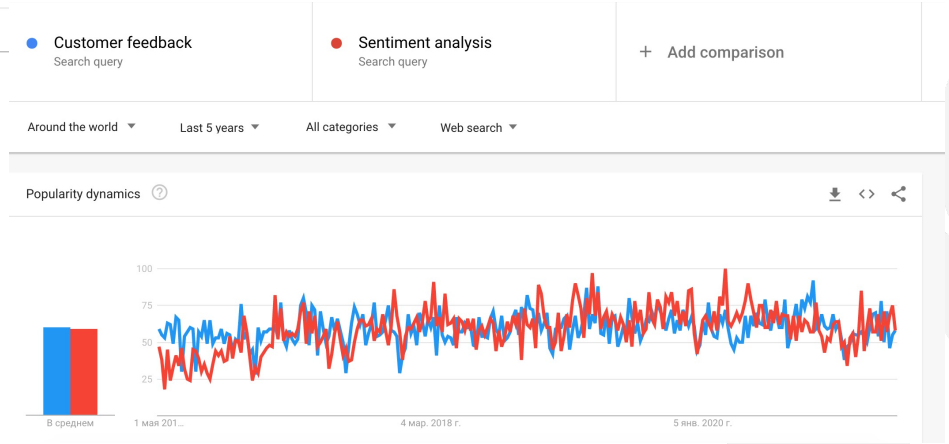
Natural Language Processing predictive modelling task.

Opinion polarity: Positive, Negative, Neural



Background





Google trends (www.google.com/trends). Relative popularity of search “sentiment analysis” & “customer feedback”

NAZARBAYEV UNIVERSITY

Background

Search “Sentiment analysis” on ScienceDirect (https://www.sciencedirect.com/) gave us

58 450 results

- 2020 year – 5627 papers
- 2021 year – 3474 papers

Sentiment analysis - one of the growing research areas in Computer Science

NAZARBAYEV UNIVERSITY

Motivation

Anne's Favorite @Frost_Sinatra · 15m
Yo @SouthwestAir somebody turned off my drink coupons can U help me??

Southwest Airlines @SouthwestAir
Follow

Replying to @Frost_Sinatra

Oh, no! If you can send us a DM with your Rapid Rewards nu
take a look into yo

Send a private message

Amazon Help @AmazonHelp · 12h
@NileshM1432 We get your **concern**. Kindly follow our Twitter page and **DM us**. We'll assist you further. ^BS

3

Samsung Support US @SamsungSupport · Oct 9
Replying to @EntGoldenchild
Hello there. Welcome to our Social Media Support Cha
providing me with more details about your Samsung d
DM, to have more room for conversation? ^Yasme

Orna McCollum doesn't recommend **Stitch Fix**.
1 August at 17:57 ·

I received my box today and the stylist did not read my style notes. Everything in the box was not my style, it's all being returned and I have turned off receiving anymore fixes. I am very disappointed that I cannot cancel my account but I have cancelled my payment method. Definitely will not be recommending this subscription.

1 comment

Like Comment Share

Veronica Hawk reviewed **SizzorS salon** — 1 ★
March 3, 2015 ·

I called twice to reschedule my appointment. No one answered. I went with both children. Keep in mind my daughter was a client here. They tell me "no child policy" after I arrived with my two children. Double standards! It's okay to bring kids as long as you pay them, but since I was getting my hair done today and not my daughter the children weren't welcomed. This place is a joke and extremely rude.

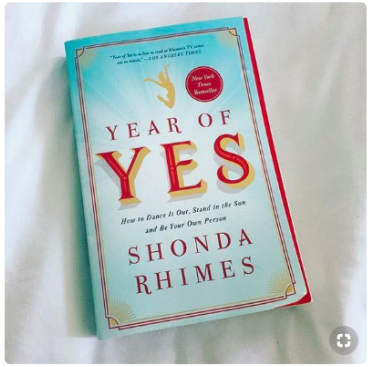
Like Comment Share

1

SizzorS salon Thank you for giving us the opportunity to address both of the issues in your review. We do indeed have a 'Children Policy' which is posted on our website and at the front desk of the salon. I have included it here: "We love children, some of us eve... See More
March 5, 2015 at 7:05pm · Like · 3

Send Select Save

instagram.com



Photos and Comments

Photos Comments

Tried this Pin? Add a photo to show how it went

Add photo

Laura Bradbury, Writer saved to **Currently Reading**
Just finished this fascinating book by Shonda Rhimes - mastermind creator and writer of such obscure TV shows as Gray's Anatomy and Scandal. I am of two minds regarding this telling of how Rhimes' life was transformed when she made a vow to start saying "yes" in her life when previously she would always say no. I loved the content and Rhimes' honesty about motherhood, career success, and fear. The change in her and her life was truly extraordinary. On the other hand I disliked her prose style to **Less**

t honed in
s take us a
get to know
u'd still
s at
am will be

NAZARBAYEV UNIVERSITY

Motivation

Social Networks

Online communication. Data flow. Recommendation system

problems

- Difficult and time-consuming to filter the information
- Relationships heavily rely on correct interpretations
- Demand for getting valuable data, optimization of the whole mechanism of the text analysis

NAZARBAYEV UNIVERSITY

Thesis objectives

- Introduce methods for sentiment analysis
- Compare different approaches [Machine Learning, Rule-based, Statistical]
- Explore the effectiveness of pre-trained models
- Investigate the application of methods

NAZARBAYEV UNIVERSITY

Modern approaches

BERT (Bidirectional Encoder Representations from Transformers) is state of the art for wide range of tasks in NLP

Software & Tools:

- Google Analytics & Alerts
- Tweet Statics
- Social Mention
- Marketing Grader

NAZARBAYEV UNIVERSITY

Literature Review

- Machine Learning
  - Supervised
  - Unsupervised
- Lexicon – based
  - Dictionary - based
  - Corpus - based
    - Statistical

NAZARBAYEV UNIVERSITY

Literature Review

Introduction and use of sentiment analysis algorithms, how they are implemented, what kind of architecture is used.

Kazakh language – shows 60% of accuracy
Kazakh language is used with Russian language – shows over 70% of accuracy

Problems & Limitations during sentiment analysis:

- Mistakes in words
- Unstructured text
- Lack of labeled learning examples
- Many facets of the language are not taken into account, as negation
- Subjectivity issue

NAZARBAYEV UNIVERSITY

Methodology: algorithms & techniques overview

- Logistic regression – observation into one of two classes
- LSTM – one of the widely used and studied methods. Over 3000 papers found from (https://paperswithcode.com/), multiple times bigger than other method related papers and takes 2nd place after time series papers.
- BERT – state of the art language model for NLP
- Polyglot – offers wide range of analysis and board language coverage supporting 136 languages for sentiment analysis task
- TextBlob – widely used library for NLP tasks, including sentiment analysis

NAZARBAYEV UNIVERSITY

Machine Learning approach

Logistic Regression

Discriminative and feature based model, does predictive analysis for classification problem.

Linear algorithm with a non-linear transform on output.

Classifier extracts set of weighted features from the inputs, takes logs, combines them linearly.

Long-Short Term Memory

Part of Recurrent Neural Networks, learn long-term dependencies. It has memory blocks, each containing and input and output gate.

BERT

Makes use of Transformers, an attention mechanism that learns contextual relations between words in a document

Bidirectional
Reads the entire sequence of words as once, allows the model to learn the context of a word based on all of its surroundings

Transformers
- encoder: reads the text input
- decoder: produces a prediction

NAZARBAYEV UNIVERSITY

Statistical approach

Polyglot

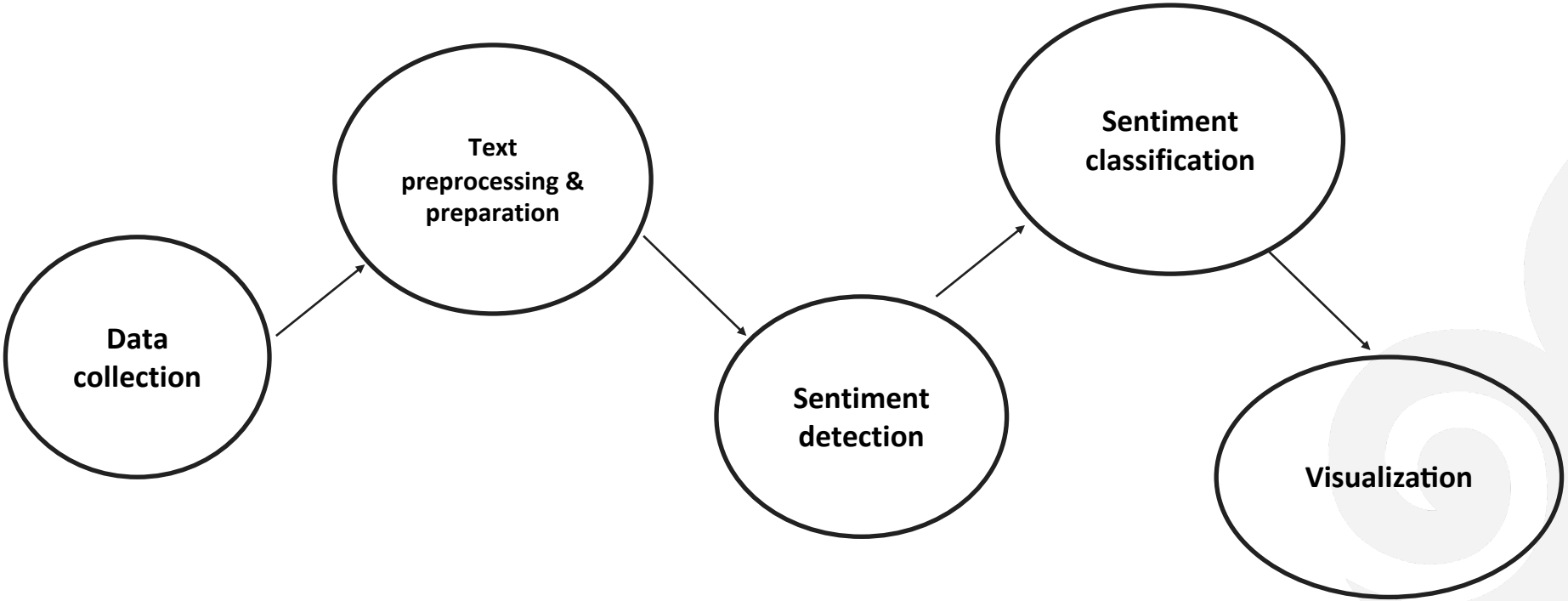
Unigram modeling approach
Polarity lexicons for 136 languages

TextBlob

Focused on pattern analyzer, returns polarity and subjectivity values

NAZARBAYEV UNIVERSITY

Methodology: Experiment work



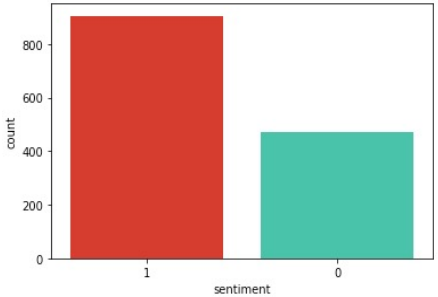
NAZARBAYEV UNIVERSITY

Data collection

Open-source dataset – Sentiment140 from Twitter

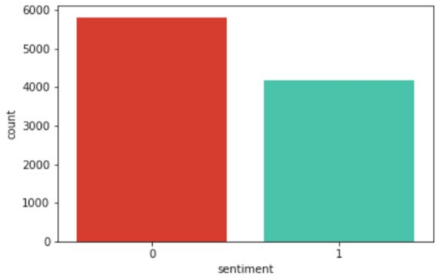
Dataset 1 - 1 700 lines

- Columns: Index, Sentence, Sentiment, Polarity, Sentiment type
- Target class balance: Positive – 0.658 (1120), Negative - 0.342 (580)



Dataset 2 - 10 000 lines

- Columns: ItemId, Sentiment, Sentence
- Target class balance: Positive – 0.4188 (4200), Negative - 0.5812 (5800)



NAZARBAYEV UNIVERSITY

Text preprocessing

Data cleaning:

- Removing the number of white space characters
- Removing the underscore
- Removing the symbols
- Removing white spaces
- Removing digits
- Removing empty lines
- Removing tags

	sentence	sentiment
0	is so sad for my APL friend.....	0
1	I missed the New Moon trailer...	0
2	omg its already 7:30 :O	1
3	.. Omgaga. Im sooo im gunna CRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...	0
4	i think mi bf is cheating on me!!! T_T	0
5	or i just worry too much?	0
6	Juuuuuuuuuuuuuuuuussssst Chillin!!	1
7	Sunny Again Work Tomorrow :-l TV Tonight	0
8	handed in my uniform today . i miss you already	1
9	hmmmm.... i wonder how she my number @-)	1

	sentence	sentiment
0	is so sad for my apl friend	0
1	i missed the new moon trailer	0
2	omg its already o	1
3	omgaga im sooo im gunna cry i ve been at this dentist since i was suposed just get a crown put on mins	0
4	i think mi bf is cheating on me t t	0
5	or i just worry too much	0
6	juuuuuuuuuuuuuuuuussssst chillin	1
7	sunny again work tomorrow tv tonight	0
8	handed in my uniform today i miss you already	1
9	hmmmm i wonder how she my number	1

Sentiment Selection - Extracting features

Finding valuable data that contains more information

CountVectorizer – converts collection of text documents to a matrix of token counts.

```
[[   0    0    0 ...   12    5  226]
 [   0    0    0 ...   77  876 1247]
 [   0    0    0 ...   86  189  393]
 ...
 [   0    0    0 ...  108   49   43]
 [   0    0    0 ...    3  330   10]
 [   0    0    0 ...    9    3 1486]]
```

NAZARBAYEV UNIVERSITY

Model building

Logistic regression

- scikit-learn (https://scikit-learn.org/)

LSTM

- keras (https://keras.io/)

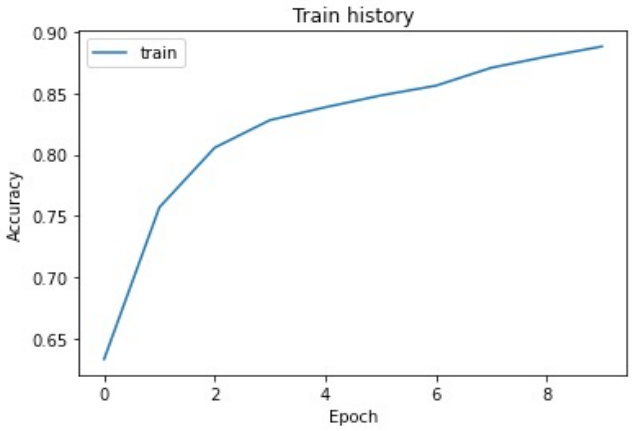
BERT

- Pytorch (https://pytorch.org/)

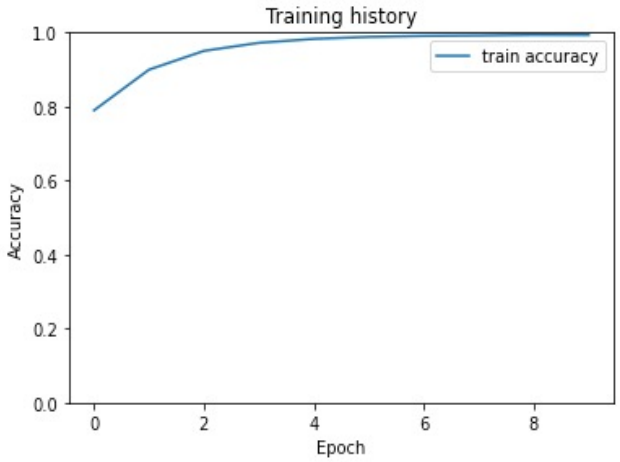
Number of epochs – 10

NAZARBAYEV UNIVERSITY

Training



Train accuracy history of LSTM model



Train accuracy history of BERT model

LSTM – 0.82, Logistic Regression – 0.77, BERT – 0.81

NAZARBAYEV UNIVERSITY

Polyglot

Detecting “Positive”

sunny again work tomorrow tv tonight	
sunny	0
again	0
work	1
tomorrow	0
tv	0
tonight	0
Positive	

hmmmm i wonder how she my number	
hmmmm	0
i	0
wonder	1
how	0
she	0
my	0
number	0
Positive	

lt this is the way i feel right now	
lt	0
this	0
is	0
the	0
way	0
i	0
feel	0
right	1
now	0
Positive	

Detecting “Negative”

i missed the new moon trailer	
i	0
missed	-1
the	0
new	0
moon	0
trailer	0
Negative	

or i just worry too much	
or	0
i	0
just	0
worry	-1
too	0
much	0
Negative	

is so sad for my apl friend	
is	0
so	0
sad	-1
for	0
my	0
apl	0
friend	0
Negative	

NAZARBAYEV UNIVERSITY

Sentiment Classification - Testing

10 random sentences using Document Generator Library (https://pypi.org/project/essential-generators/)

```
1 Persons. The the lower-density surface zone is known as the length and movement
2 Density zones: two existing customs unions: Mercosur and the Mediterranean trade.
3 Owls, Carolina with sporadic rainfall while parts of
4 Midtown, and a move into the ground in what is right. Evil or bad
5 Italian sausage. than a place name.
6 Physician Asaph downdrafts within the Boreal Kingdom and Empire), and the Arabian
7 XML dialect. entrance to
8 And testified colloquial use of effect size statistics, rather than the speed of light in
9 English languages. explain properties of the
10 Ten floors has rather warm summers, with a salad
```

NAZARBAYEV UNIVERSITY

Sentiment Classification - Testing

1. Persons. The the lower-density surface zone is known as the length and movement

[BERT – 0, LSTM – 1, Logistic Regression – 1, Polyglot – 1 (neural)]

2. Density zones: two existing customs unions: Mercosur and the Mediterranean trade.

[BERT – 0, LSTM – 1, Logistic Regression – 0, Polyglot – 1 (neural)]

3. Midtown, and a move into the ground in what is right. Evil or bad

[BERT – 0, LSTM – 0, Logistic Regression – 0, Polyglot - 1 (neural)]

4. Ten floors has rather warm summers, with a salad

[BERT – 1, LSTM – 0, Logistic Regression – 0, Polyglot – 1]

NAZARBAYEV UNIVERSITY

Evaluations

TP – model predicted the actual value correctly and it shows a positive result
TN – model predicted the actual value correctly and it shows a negative result
FP – model predicted the actual value to be positive and it is incorrect
FN – model predicted the actual value to be negative and it is incorrect

- Accuracy (TP + TN) / Total – proportions of correct predictions
- Precision TP / (TP + FP) – how many values are predicted correctly
- Recall TP / (TP + FN) – how many actual values predicted correctly
- F score – weighted method of precision and recall
- Confusion matrix

NAZARBAYEV UNIVERSITY

Results

Evaluation metrics	**LSTM**	**Logistic regression**	**BERT**
Accuracy	0.82	0.77	0.81
Precision	0.74	0.76	Negative – 0.7 Positive – 0.61
Recall	0.74	0.64	Positive - 0.81 Negative – 0.79
F score	0.74	0.67	0.82

NAZARBAYEV UNIVERSITY

Results

Passed for Logistic regression 5 positive and 5 negative sentences

	Sentence	Predictions	Expect	Results
0	feeling strangely fine now i m gonna go listen to some semisonic to celebrate	False	True	False Negative
1	handed in my uniform today i miss you already	False	True	False Negative
2	you re the only one who can see this cause no one else is following me this is for you because you re pretty awesome	False	True	False Negative
3	uploading pictures on friendster	False	True	False Negative
4	thanks to all the haters up in my face all day	False	True	False Negative
5	this weekend has sucked so far	False	False	False Positive
6	just worry too much	False	False	False Positive
7	i missed the new moon trailer	False	False	False Positive
8	is so sad for my apl friend	False	False	False Positive
9	isnt showing in australia any more	False	False	False Positive

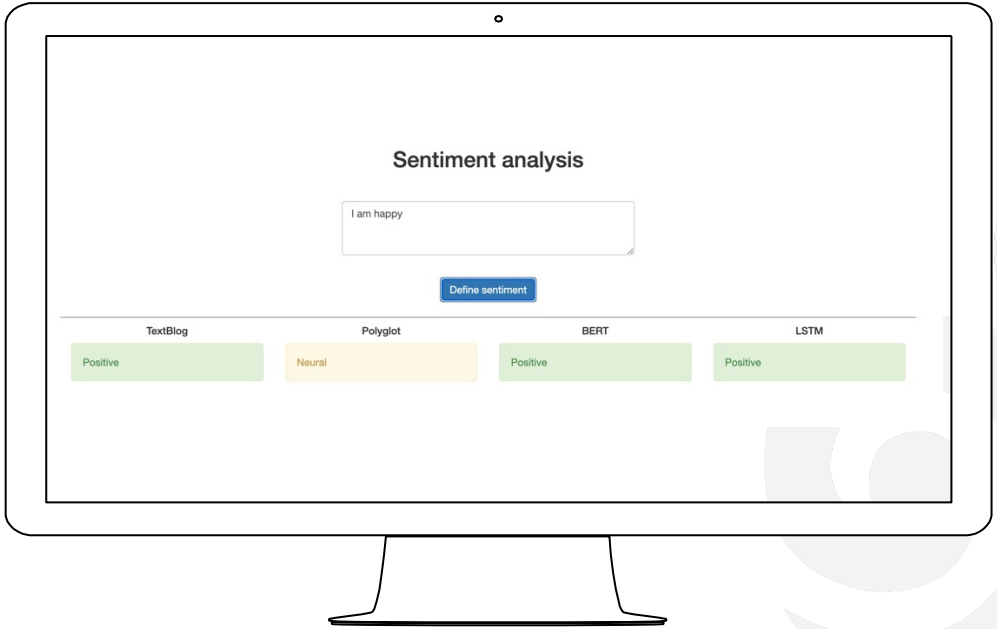
Looked to the results of others that did sentiment analysis with this open-source dataset and found that their Logistic regression model hit accuracy of 0.82 [Kritika Rupauliha's solution from Github (https://github.com/rkritika1508/Sentiment-Analysis/blob/master/Fifth.ipynb)].

NAZARBAYEV UNIVERSITY

Data visualization

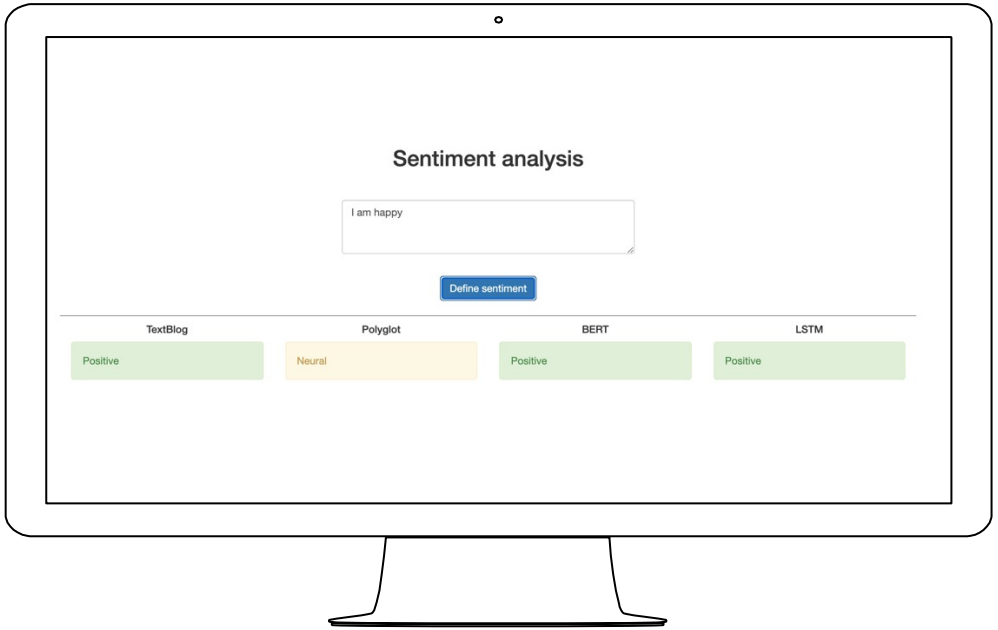
Web application:
- Flask framework
- JavaScript + Python

Based on following methods:
- LSTM
- BERT
- Polyglot
- TextBlob



NAZARBAYEV UNIVERSITY

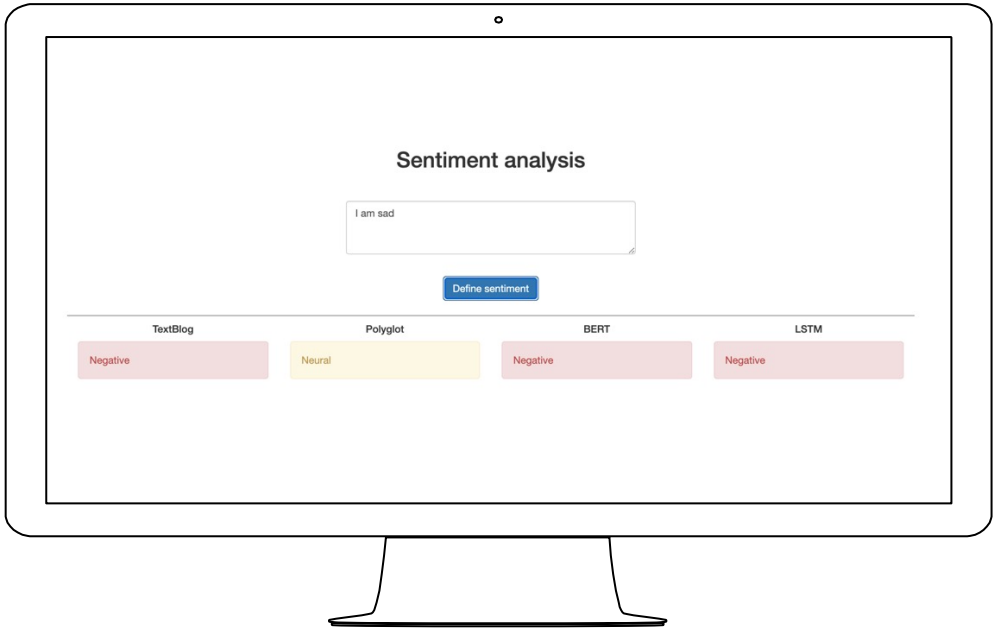
Data visualization



Web interface
Positive and Neural sentiments

NAZARBAYEV UNIVERSITY

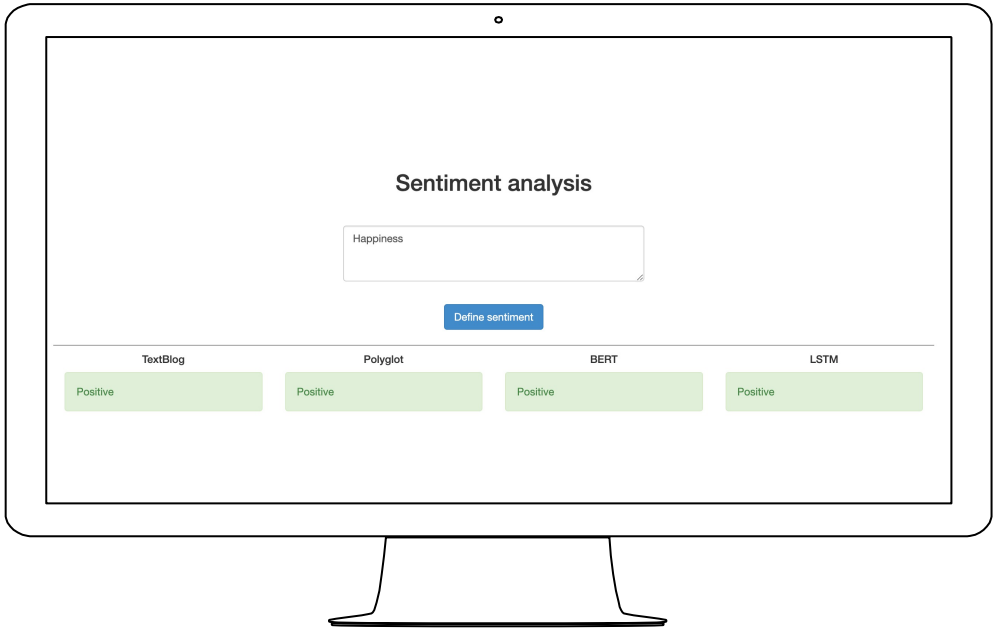
Data visualization



Web interface
Negative and Neural sentiments

NAZARBAYEV UNIVERSITY

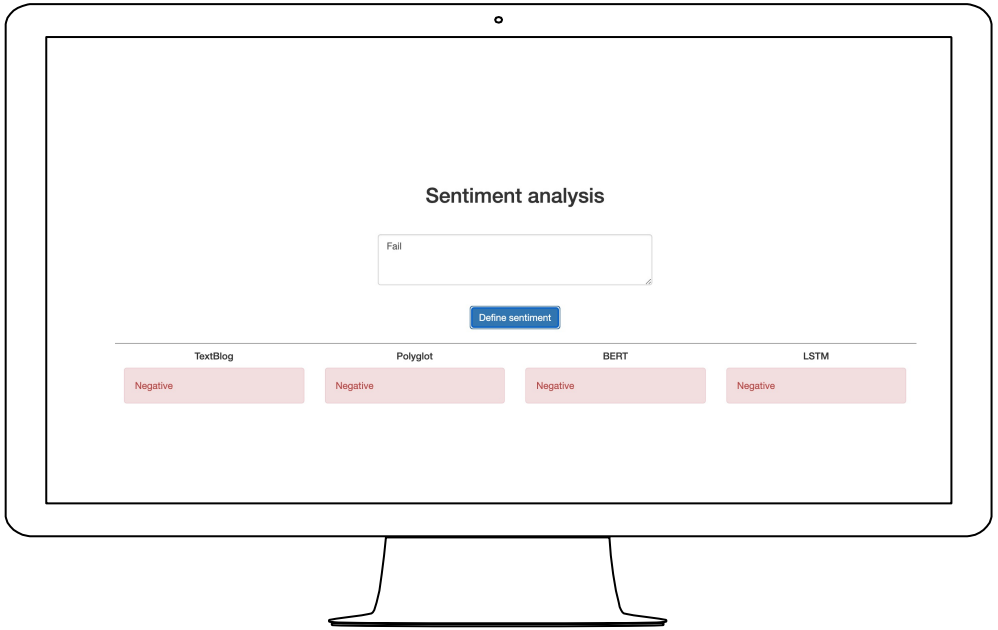
Data visualization



Web interface
Positive sentiments

NAZARBAYEV UNIVERSITY

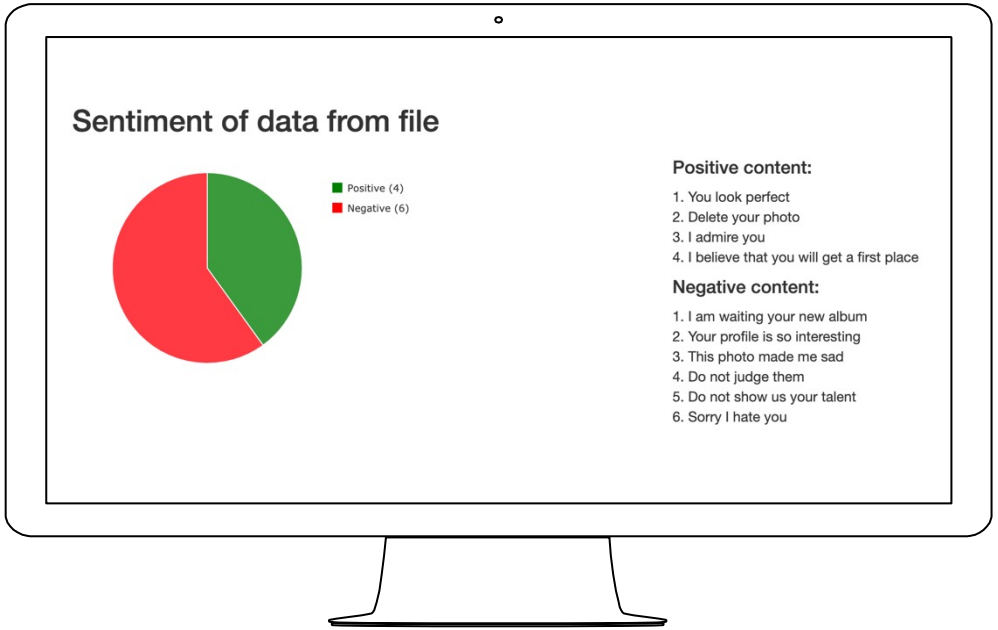
Data visualization



Web interface
Negative sentiments

NAZARBAYEV UNIVERSITY

Data visualization



Web interface
Sentiment analysis chart

NAZARBAYEV UNIVERSITY

Discussion

Sentiment prediction from different methods represented different values

1

Reason: Class distribution was imbalanced

Improvement: Truncate & pad input sequences,
use class weighted loss function, up sample class sharing

2

Reason: Targeted only “Positive” & ”Negative” classes

Improvement: Take into consideration “Neural” class

NAZARBAYEV UNIVERSITY

Discussion

Sentiment prediction from different methods represented different values

3

Reason: Did not consider sarcastic sentences, negations and did not perform stemming process

Improvement: Use Natural Language Tool Kit

4

Reason: Domain of the tested sentences

Improvement: Use social networks API

NAZARBAYEV UNIVERSITY

Application

- Businesses are interested in what customers are saying about their products, wants their brand perceived positively ⇨ customer support, needs, experience, product analysis, branch health, competitive research

- Individuals are interested in getting / reading positive materials ⇨ filtering news, so that person can read only good news

- Individuals are interested in what others are saying about their personality, job and performance ⇨ configuring social networks so that it filters comments or removes negative ones

NAZARBAYEV UNIVERSITY

Conclusions

Sentiment analysis is one of the most popular tasks in text classification

LSTM – popular & simple to implement

Logistic regression - can be advantageous if there is low dimensional data, shows better results if there is a large dataset.

BERT – requires more computational power and time.

Polyglot & TextBlob – does not need prior training, can show result quickly, because of its execution time. However, they have no learning competence. TextBlob is likely failing on large sentences. Do not take into account how words are combined in a sequence.

NAZARBAYEV UNIVERSITY

Conclusions

BERT pre-trained model & ready libraries give more precise results.

⇩

Machine Learning algorithms may be advantageous for a specific sentiment analysis task, while pre-trained models & libraries can be applied to multiple domains & languages.

Social media users use multiple languages to express their opinion.

⇩

Impossible to analyze data without error, need to take into consideration all the features of the language, improve preprocessing and do experiments on large datasets

NAZARBAYEV UNIVERSITY

Future work

- Removing words that do not contain sentiment. For example, pronouns.
- Removing repeated letters
- Handling Part of Speech and Point – Wise mutual information
- Tuning the hyperparameters. For example, Grid Search.
- Scaling the feature and normalization
- Do experiments on large datasets

NAZARBAYEV UNIVERSITY

References

- Abhilasha Tyagi, Naresh Sharma, Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic, International Journal of Engineering & Technology
- Google Trends (https://trends.google.com/)
- Nilesh Shelke, Shriniwas Deshpande, Vilas Thakare, Statistical Approach for Sentiment Analysis of Product Reviews, International Journal of Computer Science and Network, Volume 5, Issue 3, June 2016
- Medhat, W., Hassan, A. & Korashy, H, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, 5.4: 1093-1113.
- Pang B, Lee L, Opinion mining and sentiment analysis, Foundations and Trends in Information retrieval, 2, 1-135.
- Banu Yergesh, Gulmira Bekmanova, Altynbek A, Sentiment analysis of Kazakh text and their polarity, February 2019.
- Yergesh, B., Mukanova, A., Sharipbay, A., Bekmanova, G., Razakhova, B.: Semantic hy-
- per-graph based representation of nouns in the Kazakh language. Computacion y Sistemas,
- 18 (3), pp. 627-635.
- Junfei Qiu, Qihui, Guoru Ding, Yuhua Xu, Shuo Feng, A survey of machine learning for big data processing, EURASIP Journal on Advances in Signal Processing, 67 (2016), https://doi.org/10.1186/s13634-016-0355-x.

NAZARBAYEV UNIVERSITY

Thank you !