Sentiment Comparison between Native Language and English Subreddits of Countries

Xuehuai He, Pomona College

Inspiration and background

r/finland
Have you seen it, Aragorn?
The white snowy pictures of Rovaniemi. Auroras glimmering like a spike of pearl and silver
r/suomi
It is a barren wasteland, riddled with fire ash and dust. The very air you breathe is a poisonous fume

When I was browsing the Finnish language community on reddit r/Suomi (meaning Finland in Finnish), I saw this meme highlighting the contrast between the sentiments of the native language subreddit (r/Suomi) and the corresponded English subreddit (r/Finland). Although the topic of both subreddits are about the country of Finland, the contrast between the sentiment on the place from Finnish speakers (assumingly local people) and English speakers (assumingly tourists, foreigners and expats) is pretty hilarious. I am thus intrigued to see what and how big the differences are through analysing the corresponded corpora for different countries.

Methods

- Scrap the **English and native language subreddits** of different places in the world and compare sentiments.
- Train a **sentiment classifier** on one language and machine translate it into English.
- Machine translate all other corpora into English and **analyse their sentiments**.
- More on translation: first compare if the sentiment of the translated corpora is **similar enough to the original corpora** by sampling a few languages.
- Compare and contrast:
  1. **English vs Native sentiments**: which group says more good things about the place?
  2. How **the contrast is different for different countries / places around the world.**

Building the classifier

There are several ways to build a classifying model that tells which sentence is positive or negative. They all first **learn from a set of training data** that has the sentiments already labelled on each sentence. They extract **how frequently certain words appear** in positive vs. negative statements, and use them to predict the sentiments of new sentences.

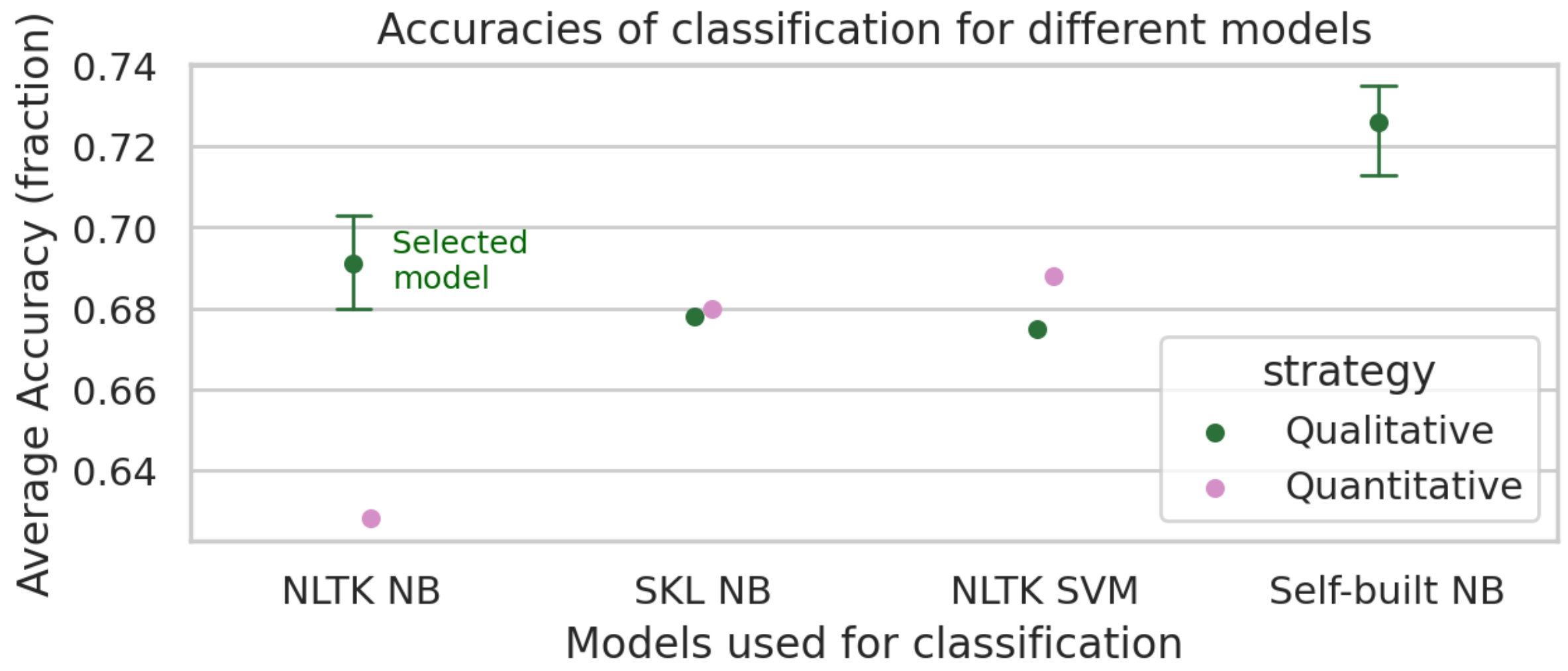
The **Naïve Bayes** (NB) method uses a simplified Bayes theorem and is computationally simpler, but would yield a lower accuracy when the size of training data is small. The **Support Vector Machines** (SVM) method computes a hyperplane that separates the data. It takes longer but works better when the size of training data is small.

In this investigaion, the training corpus is the FinnSentiment corpus produced by Turku University NLP, originally in Finnish. The corpus is **clearly annotated, recent and relevant**. The models tested are the **NLTK NB module, NLTK SVM, Scickit-learn NB, and a self-built NB.** They are trained on the same corpus with 2000 positive and 2000 negative sentences in Finnish and English, of which **randomly** 90% is used in training the model and 10% is used in testing.

Since Finnish is highly inflected, the text is **lemmatised** using the Voikko library before analysing. English sentences are not lemmatised but simply **tokenised** into individual words separated by spaces.

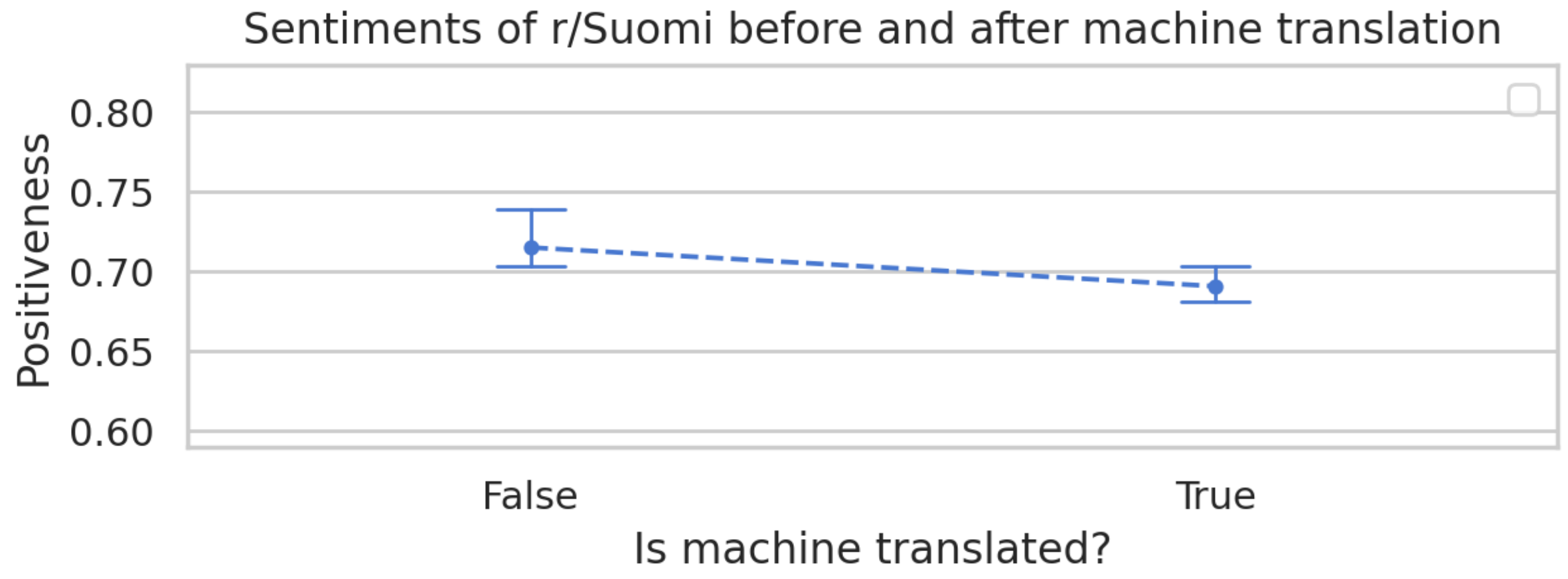
Comparison and choice of classifiers

The result after testing all classifiers is presented. The **qualitative strategy determines the sentiment based on whether certain lexicons exist or not** in the text, while the **quantitative one on counting how many times each lexicon appears in the text**.



As shown, the self-built model and the qualitative NLTK NB model yields the highest accuracy. However, further experimenting suggest that the self-built model skews negative compared with all other models. Therefore, **the qualitative NLTK NB model is chosen to analyse all texts**.

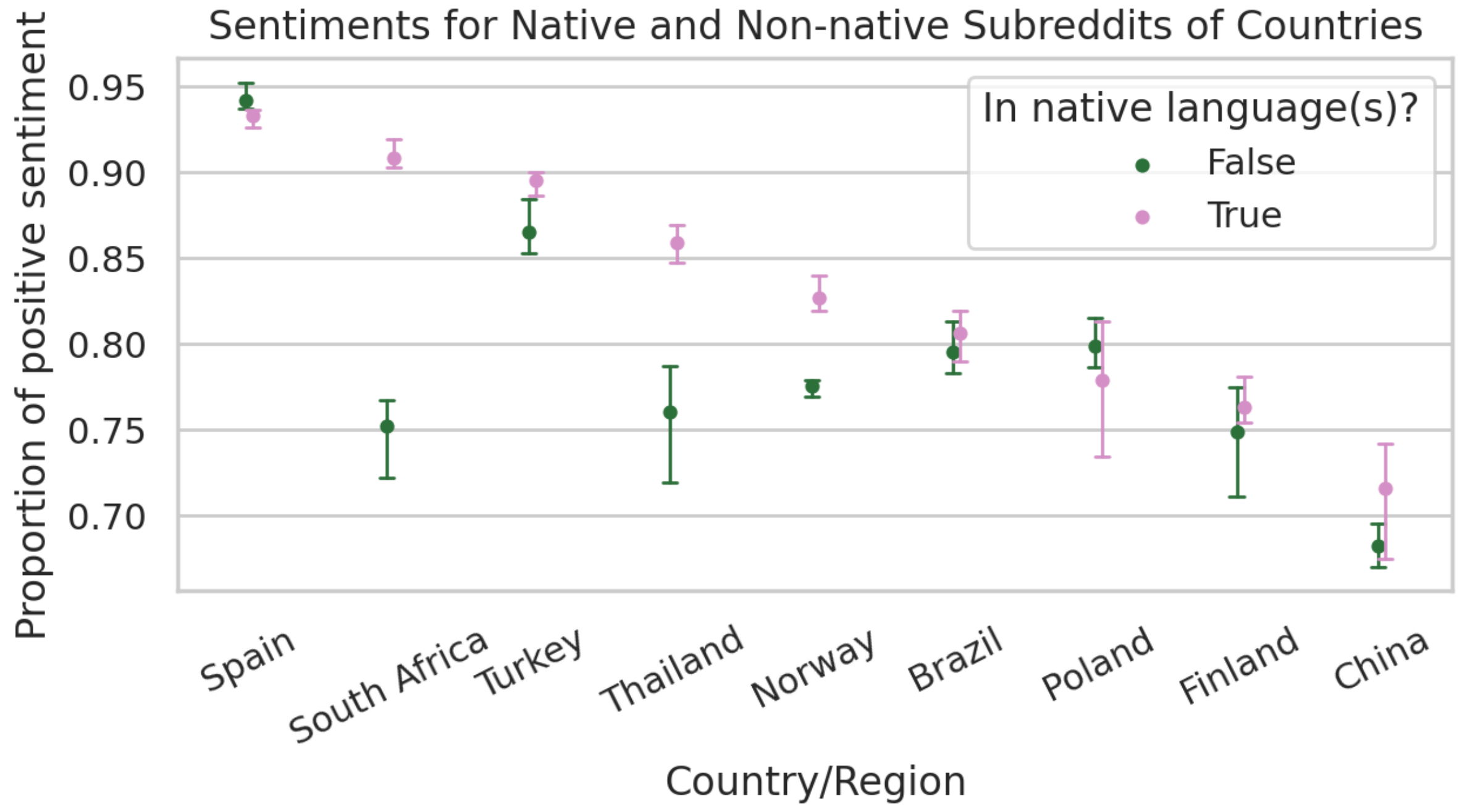
The **impact on sentiments from machine translation** is evaluated on analysing r/Suomi in Finnish and English (Google translated from Finnish):



Machine translation only altered the sentiment by 2.5% with overlaps of error bars between the original and translated sentiments. Therefore, it can be concluded that **it is safe to do sentiment analysis on machine-translated texts**.

Sentiment analysis result

As the model is prepared, **50 top posts** are scraped and translated from each of the native and English subreddits of **10 countries** on Apr. 27, 2022. The sentiment result is as follows:



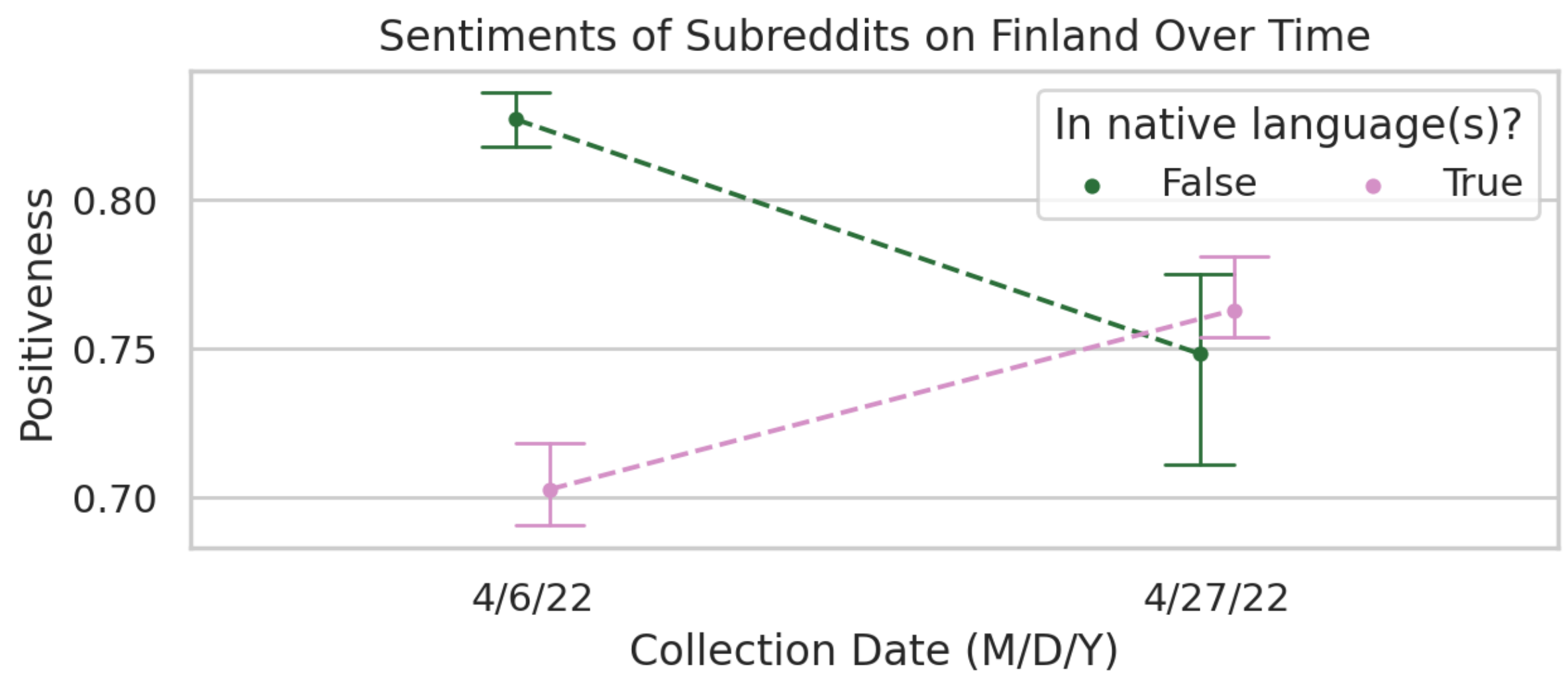
Where the green data points represent the sentiment positiveness of subreddits in English and pink ones in the native languages of the corresponding countries.

Result interpretation

One feature immediately standing out is that for most countries, **the sentiments of the native and English subreddits are, contrary to the meme, very similar** with a difference less than 5%. In addition, the native subreddits are usually **even more positive** than their English counterpart, especially with South Africa's native subreddit r/RSA 15% more positive. However, since Reddits are very **topic- and time-relevant**, some further analysis is done.

Change of sentiment over time

Reddit updates extremely rapidly according to what is happening in the world. Thus, **the sentiments would vary greatly over time**. Taking subreddits about Finland as an example, **30 top posts are scraped on Apr. 6, 2022 and 50 on Apr. 27, 2022, after 3 weeks.** The variation of sentiment is shown below:



On Apr. 6, when the meme went viral, **the sentiment in r/Finland was indeed higher than that in r/Suomi**. However, 3 weeks later, the sentiment became almost identical.

Topic reading

In addition, different subreddits might be talking about **different topics, which would influence the sentiment**. For instance, **both subreddits for Spain have a very high positiveness**. After manual reading, it is found that the topics are mainly about **food, studying and travelling**, with **few controversial topics**. For South Africa, **the native corpus has a significantly higher positiveness**. Comparative analysis found that while the native one was mainly around job seeking, the English one was on **more serious topics** like scamming, drunk-driving and medical issues.

Evaluation and further directions

One limitation of the data is that it might **lose its significance shortly after Apr. 27, 2022**; in the future, a more rigorous study could be done on a larger sample of data collected over a longer period of time. In addition, **the size and activeness of the subreddits naturally vary a lot**, which might cause the data to be skewed. Nonetheless, it provides a perspective on how natives and foreigners would perceive a place similarly or differently.

References

[1] Mikael Ahonen. Finnish stemming and lemmatization in python.

[2] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 127. Association for Computational Linguistics.

[3] Lars Borin and Forsberg. Korp – the corpus infrastructure of språkbanken.

[4] DataTechNotes. Sentiment classification with NLTK naive bayes classifier.

[5] Suhun Han. Googletrans. original-date: 2015-06-05T08:35:11Z.

[6] Tero Kemppi. Ubik sentiment. original-date: 2017-02-19T10:13:01Z.

[7] Krister Lindén and Sam Hardwick Tommi Jauhiainen. FinnSentiment, source.

[8] Bird Loper, Ewan Klein, and Steven Edward. *Natural Language Processing with Python.* O'Reilly Media Inc.

[9] MysteriousRony. Mielipide suomesta eri alaredditeissä.

[10] Harri Pitkänen, Teemu Likonen, and Flammie Pirinen. voikko/corevoikko. original-date: 2012-06-28T17:42:07Z.

[11] Jason D. M. Rennie and Ryan Rifkin. Improving multiclass text classification with the support vector machine. Accepted: 2004-10-20T21:03:52Z.