Sentiment Analysis of the CIA World Factbook

Panagiotis Podiotis
Department of International and European Studies
University of Macedonia
Thessaloniki, Greece

May 22, 2020

List of Tables

Table 1 - Sentiment predictions per model. 5

Table 2- Highest ranked (most positive) text per model. 5

Table 5 - Cumulative Sentiment per Entry (only textual columns) 6

Table 6 – Cumulative Sentiment per Entry (all original columns) 6

Table 7 - Highest and Lowest ranked entries, cross-model common occurrences. 7

Table 8 - Cumulative Sentiment per data-field (all original columns) 7

Table 9 - Sentiment per countries' relation with the US. 8

Table 10 – Various statistical measures of combined YouGov & Factbook Sentiment datasets. 8

Table 11 - t-Test p values between samples. 9

Table 12 - Mann-Whitney U Test p values. 9

Table 13 – Correlation between Sentiment and Policy for all Labels. 10

Content

List of Tables 2

Introduction 3

This Paper 3

Sentiment Analysis 4

Findings 5

Strengths & Weaknesses 11

Conclusion & Ideas for further research. 11

References 12

Introduction

Complexity Sciences (Caws, 1963), Data Mining, Machine Learning and Computational Statistics have existed for over 50 years. Such concepts existed within the private sector, technical university degrees and governmental institutions. Lately, the emergence of Big Data (Davenport, et al., 2012) and social media drew the attention of the literature (Guy, 2012) initiating the era of the “Fourth Paradigm” (Hey, et al., 2009). Data Science was brought once again to light capturing the attention of governments, the academia and the private sector. New tools and ideas like deep learning and Natural Language Processing (NLP) were developed to tackle with this new reality and enhance the existing background. Furthermore, the democratization of knowledge on the web along with the rise of commercial computation power enabled researchers of non-technical sciences to conduct multidisciplinary research on the intersect of Data Science, Computer Science and their respective field. Still, many political science theorists and practitioners remained skeptical of these new tools. Most of them were convinced after experiencing the success of the online recruiting campaigns of ISIS, the appearance of the “alt-right” internet movement, the formation of a “meme” culture and the relevancy of social media disinformation to foreign policy (Operation InfeKtion: How Russia Perfected the Art of War, 2018). These events, among others, started a quest of using Data Science tools and techniques within Political Science.

This is the point where International Relations, as a subfield of the latter (Kouskouvelis, 2007, pp. 22-23) came into play with Computational IR (Unver, 2018). Of course, such projects and approaches existed long before the last two decades but only at high-institutional levels and not at the basis of IR. The online phenomenon surrounding US President Donald Trump along with the dawn of twitter politics provided with fertile soil for the development of NLP and Data Science research in the context of Political Science. One of the most used tools in that context is Sentiment Classification, often called “Sentiment Analysis”. It is employed in order to quantify and/or quantify human emotion/sentiment from digital data.

This Paper

After having carefully considered the main arguments presented in the Introduction, I will attempt to employ Sentiment Classification for the study of International Relations though a case study of the CIA World Factbook (hereinafter Factbook) (CIA, 2019, p. About). The Factbook as a data source offers a great opportunity of research since it provides multi-domain[1] and high confidence data related to the study of International Relations. Moreover, these data come from an agency which forms part of the spearhead of US’s foreign policy, arguably the most important actor of the International System.

This exploratory paper will employ various Sentiment Analysis and statistical tools to discover how the US portrays various nations in the CIA World Factbook. Specifically, I will try to examine whether the underlying relations between the US and various states affects how they are portrayed in the CIA World Factbook. The null hypothesis can thus be expressed as: “A country’s relation with the US does not affect the sentiment expressed in the Factbook”. Focus will be set on textual data. The latest available[2] version (2018 version, uploaded on Jan 04, 2019 01:51 PM – downloaded 15/05/2020) of the CIA World Factbook which was downloaded through the official web portal was used. The dataset format of the Factbook was produced by the author in the context of his bachelor’s dissertation (Podiotis, 2020)[3].

[1] The Factbook covers 12 thematic areas for each entry (country), namely: History, People and Society, Government, Economy, Energy, Geography, Communications, Transportation, Military, Terrorism, Transnational Issues. Each Thematic area (category) consists of numerous fields providing numeric or textual data points.

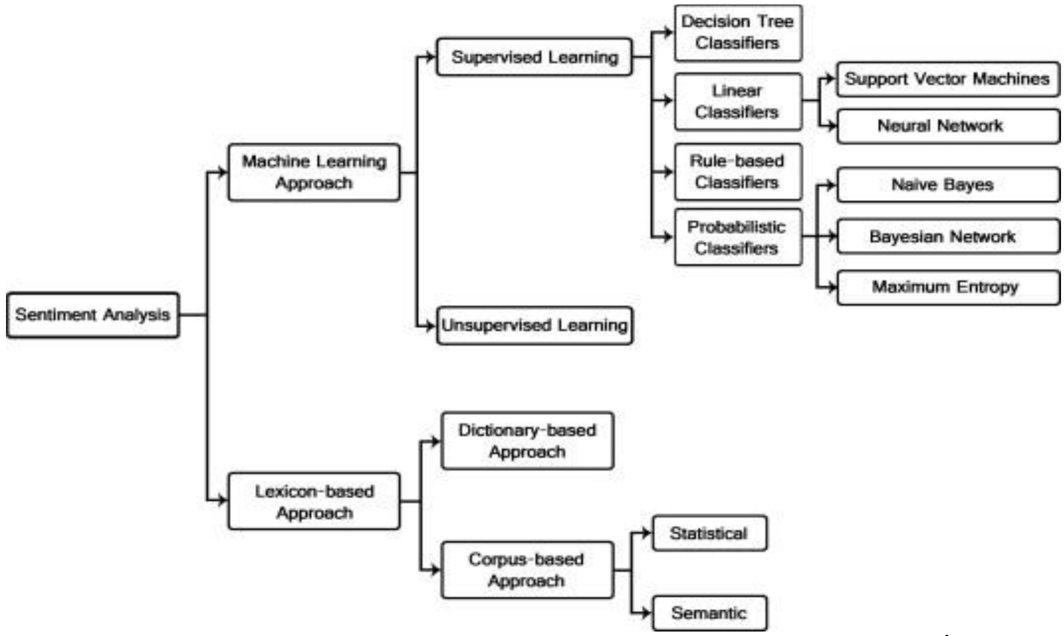
[2] There is a discrepancy between the online and the downloadable version of the CIA World Factbook with the online being the newest one. Observed by the author in his Dissertation (Podiotis, 2020).

[3] Pending title and official submission.

Sentiment Analysis

The term "Sentiment Analysis" is a widely disputed one. Many claim that this term is nothing more than a fancier way of expressing classification tasks. In an effort to avoid this debate, the term "Sentiment Classification" will be used. On a more technical level, despite this point of conflict, most agree to the below chart of approaches in conducting Sentiment Classification.

Figure 1 - Sentiment Analysis Approaches (Medhat, et al., 2015)



All of the above can be implemented by the programming language python[4], which I used for this paper. Unfortunately, due to limitations in hardware capabilities, technical knowledge, paper size and scope only few of the above were implemented. Both Lexicon-based and Machine Learning tools were used. Preprocessing and cleaning processes included: Removal of numbers, special characters, stop words, months[5], tokenization and lemmatization.

For data processing and handling, regular expressions, panda's Dataframes, numpy arrays were utilized. Statistical methods were conducted by Scipy and Pingouin. Visualization was made possible with matplotlib. For the sentiment classification task three libraries were used: NLTK (Project, NLTK, n.d.), Textblob (Loria, n.d.) and Flair (Alan Akbik, 2019). The first two quantify sentiment on a -1 to 1 scale (-1 being fully negative) while Flair binarily classifies sentiment into negative or positive while providing a confidence coefficient in a 0 to 1 scale. All outputs were exported into excel for visual inspection. All three libraries are among the most recognized and extensively used in python for natural language processing. Moreover, they are easy to use and interpret assisting non-technical degree students.

Textblob and NLTK sentiment classification algorithms are Lexicon-based and utilize "bag of words". NLTK can handle negation (ex. "bad" vs. "not bad") while Textblob cannot. Unfortunately, due to their Lexicon-based nature both libraries fail to assign sentiment to words not included in the vocabulary. Flair, on the other hand, can be considered a Supervised learning, Linear Classification Neural Network: *"Flair's sentiment classifier is based on a character-level LSTM neural network which takes sequences of letters and words into account when predicting"* (Terry-Jack, 2019). The nature of our dataset (CIA World Factbook) and the lack of labels (regarding sentiment) make it extremely unsafe and time consuming to construct new models. This is the reason existing solutions were used; in the case of Flair, a pretrained model.

[4] Version 3.8.3, using JetBrains PyCharm Community Edition 2019.2.2 x64 (JetBrains, 2000) IDE.

[5] Throughout the CIA World Factbook months along with year indicate when the various data points were updated thus increasing noise.

Findings

In total, 14768 cells containing text were found and processed, out of which the 13722 were text-only. The length of text per cell ranged from 2 up to 250 words. In total, the CIA World Factbook Corpus consists of 400.712 words (stop words, months, and other words removed during data cleaning not counted) formed by 28.645 unique words (lemmatized, words removed during data cleaning not counted).

Table 1 - Sentiment predictions per model.

	All Original Columns			Only Textual Columns		
	Positive	Negative	Neutral	Positive	Negative	Neutral
Textblob	5050	2425	7293	4817	2312	6592
NLTK	3881	2265	8622	3803	2112	7806
Flair	7323	7465	0	6744	7002	0

The "All Original Columns" approach is different to the "Only Textual Columns" because it captures small length comments accompanying numerical data.

Table 2- Highest ranked (most positive) text per model.

Model	Highest Ranked Text (most positive)	Value (-1 to 1)
NLTK	*"since late china moved closed centrally planned system more market oriented one play major global role china implemented reform gradualist fashion resulting efficiency gain contributed more tenfold increase since reform…efficient over time slight acceleration economic growth —the first uptick since —gives beijing more latitude pursue economic reform focusing financial sector deleveraging supply side structural reform agenda first announced late"*	0.9977
Textblob	*"bandar seri begawan capital boundary capital city expanded greatly increasing city area population capital increased tenfold"*	0.8
Flair	Flair performs binary classification. Texts cannot be ranked.	

Table 3 - Lowest ranked (most negative) text per model.

Model	Lowest Ranked Text (most negative)	Value (-1 to 1)
NLTK	*"current situation saudi arabia destination country men woman subjected forced labor lesser extent forced prostitution men woman south east asia middle east africa voluntarily travel saudi arabia domestic servant low skilled laborer … victim resulting unidentified victim arrested detained deported sometimes prosecuted more victim identified referred protective service previous year victim sex trafficking male trafficking victim not provided shelter remained vulnerable punishment"*	-0.9965
Textblob	*"agricultural product foodstuff wine oil product chemical product plastic rubber hide leather wood cork wood pulp paper textile material clothing footwear machinery tool base metal"*	-0.8
Flair	Flair performs binary classification. Texts cannot be ranked.	

The comparative study of Tables 2 & 3 along with the extended ranks reveals that NLTK outperformed Textblob. This is also illustrated in the instance of Great Britain. For example, the entry: “declared independence great britain recognized great britain” ranks 3rd for Textblob and 734th for NLTK. For the reasons described above, **more confidence is placed on NLTK’s sentiment assignments.**

Table 3 - Cumulative Sentiment per Entry (only textual columns)

	Most Positive (top 5)	Most Negative (top 5)
NLTK	**European Union**	Italy
	Mexico	Eritrea
	United States	Gaza Strip
	Hong Kong	Saudi Arabia
	United Kingdom	Yemen
Textblob	**Hong Kong**	Howland Island
	United Kingdom	Kosovo
	Macau	Paracel Islands
	Cook Islands	Johnston Atoll
	United States	Akrotiri
Flair (Positive=1 and Negative=-1)	Greenland	Burundi
	Guernsey	Cuba
	French Polynesia	Syria
	Faroe Islands	Libya
	Saint Pierre and Miquelon	Belarus

Bolded will be discussed below. Small island states and territories seem to be favored. Eritrea, Gaza Strip, Yemen, Cuba, Syria and Libya are all involved in conflict, suffer from unstable governance and poor living conditions. Saudi Arabia ranks low because of human rights violations which are expressed throughout the Factbook. The low placement of Italy lowers the confidence in the analysis considering that Italy is a country with high living conditions, democratic governance and generally a positive image worldwide.

Table 4 – Cumulative Sentiment per Entry (all original columns)

	Most Positive (top 5)	Most Negative (top 5)
NLTK	**European Union**	**Libya**
	Mexico	**Gaza Strip**
	Hong Kong	Zimbabwe
	United States	Saudi Arabia
	Chile	Yemen
Textblob	**Hong Kong**	Howland Island
	Haiti	Bulgaria
	Cook Islands	Paracel Islands
	United Kingdom	Johnston Atoll
	Macau	Akrotiri
Flair (Positive=1 and Negative=-1)	Greenland	**Libya**
	Faroe Islands	Belarus
	French Polynesia	Cuba
	Guernsey	**Gaza Strip**
	Wallis And Futuna	West Bank

The commentary provided for table 4 applies with the only exception being the absence of Italy from the low ranked states along with the agreement between Flair and NLTK approaches on Libya and Gaza Strip. Overall, it seems that **considering all text of the Factbook and not only of textual columns yielded better results.** Combining common occurrences in Tables 4 and 5 can help generate the table below:

Table 5 - Highest and Lowest ranked entries, cross-model common occurrences.

	Common Most Positive	Common Most Negative
All Models	**European Union**	**Libya**
	Mexico	**Gaza Strip**
	Hong Kong	
	United States	

Libya and Gaza Strip may be justified for their low ranks due to the fact that both suffer from conflict over the past years and for possessing governments not aligned with the US. On the other hand, the EU, Hong Kong and the US are justified in ranking high but Mexico is a surprise considering its high degree of corruption, migration and illicit drug trade.

Table 6 - Cumulative Sentiment per data-field (all original columns)

	Most Positive (top 5)	Most Negative (top 5)
NLTK	**government-judicial branch**	transnational-issues-illicit drugs
	government-legislative branch	transnational-issues-refugees and internally displaced persons
	government-constitution	transnational-issues-trafficking in persons
	government-flag description	geography-environment - international agreements
	government-political parties and leaders	geography-environment - current issues
Textblob	government-national anthem	government-citizenship
	economy-imports - commodities	geography-environment - current issues
	geography-environment - international agreements	government-country name
	geography-population distribution	military-and-security-military service age and obligation
	geography-climate	government-national symbol(s)
Flair (Positive=1 and Negative=-1)	**government-judicial branch**	government-political parties and leaders
	transportation-ports and terminals	government-citizenship
	people-and-society-major urban areas - population	geography-environment - current issues
	economy-imports - commodities	economy-exchange rates
	government-government type	government-country name

The accumulated sentiment per data-field is the same with Table 7 for only textual columns. On a technical note, Textblob seems to have underperformed once again by assigning very high sentiments for columns “government-national anthem” and “geography-population distribution” and extremely negative for “government-country name”. The content of the aforementioned columns was studied manually and wasn’t found to contain negative content to such an extent. Models disagree on both most positive and negative columns. Interestingly, both NLTK and Flair found “government-judicial branch” to be extremely positive. Negative-ranked fields referring to illicit drugs, environmental problems, refugees and trafficking were understandably ranked correctly.

Lastly, the distribution of Allied-Enemy states in terms of overall sentiment was studied. Establishing the relations between a nation and the US in a label-like manner is a daunting task. For this reason, the public opinion research by YouGov (YouGov, 2017) which was also used by the NYT (KATZ & QUEALY, 2017), can provide foreign policy in label format for this paper. The outcome of the survey was downloaded and countries were labeled as “Ally (numerical label 2)-Friendly (numerical label 1)-Not sure (numerical label 0 to reflect neutrality)-Unfriendly (numerical label -1)-Enemy(numerical label-2)”. The labels were then appended to the NLTK sentiments. NLTK was used due to the fact that it proved to perform the best in the context of this paper. The new Dataframe consisted of columns: Overall Sentiment and US Relations Label for each country included in the survey.

Table 7 - Sentiment per countries' relation with the US.

	Enemy	Unfriendly	Not Sure	Friendly	Ally
Total Countries	4	8	82	43	7
Positive Sentiment Countries	4	4	71	41	7
Negative Sentiment Countries	0	4	11	2	0
Mean Sentiment	2.9	-0.1	2.6	5.1	6.3

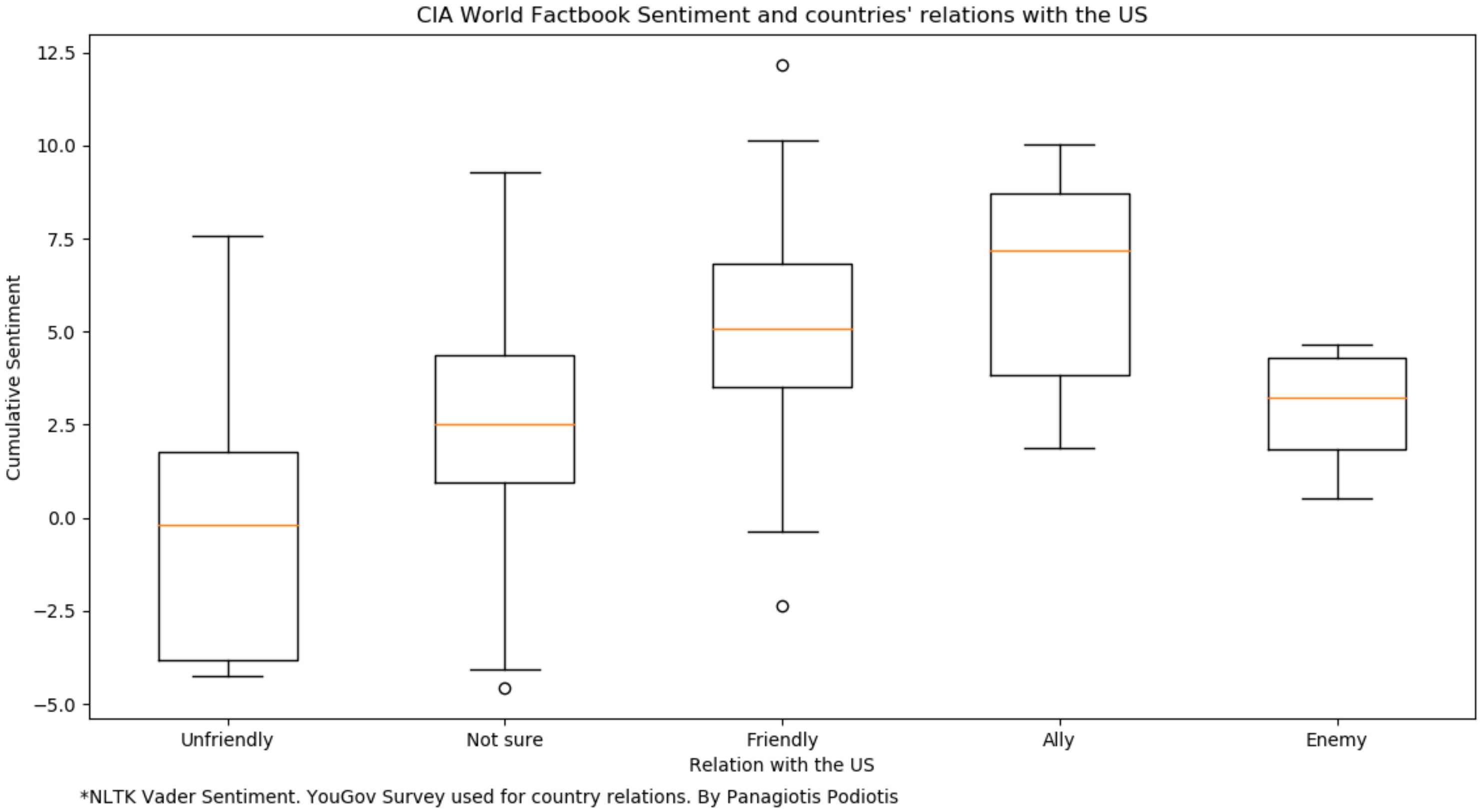
While all Enemy countries received positive sentiments, 50% of Unfriendly received negative. At the same time, all Allies received positive and only 4.65% of Friendly received a negative sentiment. The mean sentiment of each category of countries reveals something important. The mean sentiment of Allied and Friendly countries is double this of Enemies and fifty times larger than this of Unfriendly which is also surpassed by the Not Sure category!

Table 8 – Various statistical measures of combined YouGov & Factbook Sentiment datasets.

	Enemy	Unfriendly	Not Sure	Friendly	Ally
Sample Total Countries	4	8	82	43	7
Sample Size / Population* (%)	2%	4%	41%	21%	3.5%
Sample Variance	3.59	16.17	6.69	6.69	10.75
Std. Deviation	1.89	4.02	2.58	2.76	3.28
Skewness	-0.64	0.87	-0.11	-0.10	-0.57

*Population of countries considered 197 (UN members + observer states + Taiwan + Kosovo).

Figure 2 - Boxplots of different policy groups in regard to sentiment.



The relation of a country with the US is the independent variable (x) and the Factbook's sentiment regarding this country is the dependent variable (y). The population is 197 and our sample is 144. Our sample of 144 consists of five groups, described above.

Table 9 - t-Test p values between samples.

	Enemy	Unfriendly	Not Sure	Friendly	Ally
Enemy	-	0.43	0.90	0.17	0.29
Unfriendly	0.43	-	0.09	**0.02**	**0.01**
Not sure	0.90	0.09	-	**0.01**	**0.03**
Friendly	0.17	**0.02**	**0.01**	-	0.57
Ally	0.29	**0.01**	**0.03**	0.57	-

Table 10 - Mann-Whitney U Test p values.

	Enemy	Unfriendly	Not Sure	Friendly	Ally
Enemy	-	0.33	0.33	0.33	0.33
Unfriendly	0.33	-	0.02	**0.00**	**0.01**
Not sure	0.33	0.02	-	0.00	**0.02**
Friendly	0.33	**0.00**	**0.00**	-	0.35
Ally	0.33	**0.01**	**0.02**	0.35	-

From the above tests we can infer that Friendly & Ally states' sentiments are statistically similar while both are different to Unfriendly. We can safely assume that Friendly & Ally are treated differently than Unfriendly. Enemy states do not yield a sufficiently low p value (smaller than 0.05). This may be due to the small number of observations. Having establish the statistical difference between Friendly & Ally against Unfriendly, we proceed below to test correlation between Policy and Sentiment.

Table 11 – Correlation between Sentiment and Policy for all Labels.

	Sentiment	Relation with the US
Covariance	1.02	
Pearson correlation coef.	0.43	
Spearman correlation coef.	0.46	
BF10	188500	
Power	0.99	
P	<0.05	

The BF10, Power and p values increase the confidence in our sentiment findings. Unfortunately, both Spearman and Pearson coefficients are not high enough to safely establish a moderate correlation but at the same time are not low enough to claim total lack of correlation.

Finally, we can argue that:

1) The sentiment of Friendly & Ally states is statistically similar.
2) The sentiment of Friendly & Ally states is statistically different to this of Unfriendly states.

From 1 & 2 we can say that **there is a statistical difference between the sentiments of Friendly & Ally vs. Unfriendly States.**

3) A weak overall correlation (>0.4 and <0.5) exists between sentiment and country affiliation with the US. When the correlation between Sentiment-Policy Label was recalculated only for Ally, Friendly and Unfriendly States it was found to be: Pearson: 0.53, Spearman: 0.42, $p<0.05$. Pearson's coefficient surpassed 0.5 thus a weak correlation between the policy and sentiment of Allied, Friendly and Unfriendly states can be assumed.

From 1,2 and 3 it was found that there is a weak correlation between the policy and sentiment of Friendly, Allied and Unfriendly states. Friendly and Allied states are portrayed more favorably than Unfriendly.

Strengths & Weaknesses

Sentiment classification which forms the main axis of this paper has a plethora of inherent weaknesses. The most important is the high level of subjectivity. Even between human interactions sentiment can often be misinterpreted. Things get even more complicated when sentiment has to be quantified. Contrary to regression problems, sentiment classification algorithms cannot be tested and evaluated upon actual numerical values but rather on human generated labels. The complexity of human language, idioms, slang language, negation and irony can all lead to misinterpretations.

In the case of the CIA World Factbook, the following factors contributed to the underperformance of applications:

1. The dataset does not have sentiment labels thus model performance cannot be measured effectively.
2. The content is highly specialized, relevant vocabulary is used.
3. Formal language is used and sentiment-defining words like “good”-“bad” are not used often.
4. The corpus is structured arbitrarily with single-sentence comments accompanying numerical data. Text Generally is dispersed between numerous fields lacking paragraph-sentence structure.

The aforementioned underperformance of models was exemplified by the different outputs of models but was also validated by comparing assigned sentiment scores with the original text, a process which is by itself subjective.

Even though the YouGov survey was conducted one year prior (2017) to the latest available version (2018) of the Factbook, and expressed public opinion rather expert opinion, it still managed to provide with accurate labeling of the included states’ foreign relations with the US.

The CIA World Factbook contains data of high accuracy expressed by one of the most important and prestigious US governmental agency, heavily involved in International Relations.

Lastly, proven and robust libraries were used and findings were further analyzed with some of the most basic and recognized statistical methods.

Conclusion & Ideas for further research.

Even though **sentiment classification models didn’t perform optimal, statistical analysis of outcomes suggested that there is a weak correlation between a country’s affiliation with the US and the sentiment expressed regarding that country in the CIA World Factbook. This applies to countries labeled as Unfriendly, Friendly and Ally; this rule cannot be generalized to Enemy and Not sure states.**

More work needs to be conducted on the field of Data Science and International Relations. The lack of related dictionaries, datasets and labeling[6] make Data Science tasks for pure IR students difficult. The creation of multidisciplinary teams within universities, research institutes, NGOs and governmental bodies are overdue steps.

A future and extended version of this paper could study the sentiment of older Factbook versions in a time-series manner in parallel with the US’ foreign policy at the respective time point. This way, it can be studied with higher accuracy whether the sentiment of the Factbook evolves to favor allied nations in a historic manner.

[6] Various CIA World Factbook versions could be joined to created a large textual corpus which can in-turn be labeled to provide a basis for future research and development of sentiment classification tasks on foreign policy data.

References

Alan Akbik, T. B. D. B. K. R. S. S. R. V., 2019. *FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP.* Minneapolis, Minnesota, Proceedings of NAACL-HLT 2019: Demonstrations.

Caws, P., 1963. Science, Computers, and the Complexity of Nature. *Philosophy of Science*, pp. 158-164.

CIA, 2019. *Publications: Download.* [Online]
Available at: https://www.cia.gov/library/publications/resources/the-world-factbook/index.html
[Accessed 25 August 2019].

Davenport, T., Barth, P. & Bean, R., 2012. How 'Big Data' is Different. *MIT Sloan Management Review vol 54 no.1*, Fall, pp. 22-24.

Guy, R., 2012. THE USE OF SOCIAL MEDIA FOR ACADEMIC PRACTICE: A REVIEW OF LITERATURE. *Kentucky Journal of Higher Education Policy and Practice: Vol. 1 : Iss. 2 , Article 7*, Juky, pp. 1-20.

Hey, T., Tansley, S. & Tolle, K., 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond: Microsoft Research.

JetBrains, s., 2000. *JetBrains: Pycharm.* [Online]
Available at: https://www.jetbrains.com/pycharm/
[Accessed 29 August 2019].

KATZ, J. & QUEALY, K., 2017. Which Country Is America's Strongest Ally? For Republicans, It's Australia. *New york Times*, 3 February.

Kouskouvelis, I., 2007. *Εισαγωγή στις Διεθνείς Σχέσεις.* Athens: POIOTHTA.

Loria, S., n.d. *TextBlob.* [Online]
Available at: https://textblob.readthedocs.io/en/dev/
[Accessed 18 May 2020].

Medhat, W., Hassan, A. & Korashy, H., 2015 . Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal issue 5*, pp. 1093-1113.

Operation InfeKtion: How Russia Perfected the Art of War. 2018. [Film] US: New York Times.

Podiotis, P., 2020. *Towards Data Science, the CIA World Factbook.* Thessaloniki: *Pending Publication. * *Pending title and official submission.*

Project, NLTK, n.d. *NLTK.* [Online]
Available at: https://www.nltk.org/index.html
[Accessed 18 May 2020].

Terry-Jack, M., 2019. *Medium.* [Online]
Available at: https://medium.com/@b.terryjack/nlp-pre-trained-sentiment-analysis-1eb52a9d742c
[Accessed 18 May 2020].

Unver, H. A., 2018. *Computational International Relations. What Can Programming, Coding and Internet Research Do for the Discipline?,* s.l.: s.n.

YouGov, 2017. *America's Friends and Enemies.* US: YouGov.

Zalando Research, 2018. *GitHub Repository.* [Online]
Available at: https://github.com/flairNLP/flair
[Accessed 18 May 2020].