

Article

Methods for Mid-Term Forecasting of Crop Export and Production

Dmitry Devyatkin * and Yulia Otmakhova 

Federal Research Center “Computer Science and Control” RAS, 119333 Moscow, Russia; otmakhovajs@yandex.ru

* Correspondence: devyatkin@isa.ru

Abstract: A vast number of studies are devoted to the short-term forecasting of agricultural production and market. However, those results are more helpful for market traders than producers and agricultural policy regulators because any structural change in that field requires a while to be implemented. The mid and long-term predictions (from one year and more) of production and market demand seem more helpful. However, this problem requires considering long-term dependencies between various features. The most natural way of analyzing all those features together is with deep neural networks. The paper presents neural network models for mid-term forecasting of crop production and export, which considers heterogeneous features such as trade flows, production levels, macroeconomic indicators, fuel pricing, and vegetation indexes. They also utilize text-mining to assess changes in the news flow related to the state agricultural policy, sanctions, and the context in the local and international food markets. We collected and combined data from various local and international providers such as UN FAOSTAT, UN Comtrade, social media, the International Monetary Fund for 15 of the world’s top wheat exporters. The experiments show that the proposed models with additive regularization can accurately predict grain export and production levels. We also confirmed that vegetation indexes and fuel prices are crucial for export prediction. Still, the fuel prices seem to be more important for predicting production than the NDVI indexes from past observations.

Keywords: wheat production and export forecast; transformer; recurrent network; NARX; regularization; sentiment analysis



Citation: Devyatkin, D.; Otmakhova, Y. Methods for Mid-Term Forecasting of Crop Export and Production. *Appl. Sci.* **2021**, *11*, 10973. <https://doi.org/10.3390/app112210973>

Academic Editor: José Miguel Molina Martínez

Received: 29 September 2021
Accepted: 12 November 2021
Published: 19 November 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The COVID-19 pandemic triggered a drop in all types of production, a decrease in trade flows, and rupture of cross-country communications around the world [1]. On the one hand, those challenges require urgent measures to develop agriculture successfully. On the other hand, commercial food has every chance of becoming the primary export item and an alternative even to energy sources, providing international weight and social stability within many countries. At the same time, the current level of grain production does not correspond to the potential of soil resources, is not an indicator of sustainability, and the recent modern agricultural technologies are not adequately implemented. Therefore, reaching the competitive level of grain yield is the primary goal of agriculture.

Because the state resources are limited, the efforts should be focused on a small set of prospective development directions. Namely, the development of grain production and export implies discovering a restricted set of the new prospective commodity items, which would be the primary growth drivers. Moreover, that set should be revealed as soon as possible because all state-level measures in agriculture usually require several years to be implemented. This problem can be defined as a mid-term forecast of a particular commodity’s export and production level; therefore, one can analyze various export items and leave only the most prospective ones. Thus, such forecast models can be applied as a part of the software to optimize country and regional-level agricultural management

in case of insufficient resources (water regime of soils, plant nutrition, yield forecasts) and achieve better sustainability of agriculture. That problem is not trivial because of the following issues:

- Variable character of agricultural trade flows.
- Too many features influence trade flows and production. Using them all would lead to over-complex prediction models, which are not trainable with the dataset. Mid-term forecasting requires the use of complex models that consider many features and parameters from past observations, but the size of the training dataset is strictly limited [2]. Therefore, complex models can be easily over-fitted and, in some cases, provide incorrect results on unseen data [3].
- Political decisions and economic sanctions strongly affect trade flows and production, but they can hardly ever be predicted using only statistic databases. Additional information such as news and social media messages should be considered.

The paper presents neural network models for mid-term forecasting wheat production and export, considering heterogeneous temporal and spatial features such as trade flows, production levels, macroeconomic indicators, fuel pricing, and vegetation indexes. We also use text-mining to detect events in the news flow related to the state of agricultural policy, sanctions, and the context in the local and international agricultural markets. The proposed models have value in practical applications since they provide accurate mid-term forecasting of crop production and export and consider non-structural features, which can significantly affect the markets.

In this paper, we tackle the problem by answering the following research questions:

1. Can the textual features improve export and production forecasting?
2. Which neural network architectures can integrate heterogeneous structured and textual features to provide accurate forecasting?
3. Is there any training regularization to increase the forecasting reliability and reduce the overfitting?

2. Related Work

Multilayer neural network models are widely used for forecasting crop export and prediction. We believe this is because those models can naturally integrate heterogeneous market [4], production, and climate features. Bhojani et al. [5] applied a multilayer feedforward network model to forecast the wheat crop yield at the district level. They focused mainly on revealing the activation function, which would achieve the best accuracy of prediction. In this study, they evaluated various activation functions and recommended some exponential activation functions to improve the accuracy of the yield forecasting models. Experiments show that proposed activation functions are better than the 'sigmoid' activation function for agriculture datasets. Past values of various features such as weather conditions, fuel, and fertilizer prices can significantly affect the yield and trade activity. Therefore, time-series prediction models, which can consider those past observations, seem to be the most prospective. For example, Ref. [6] presents a regression model containing a convolutional neural network and recurrent neural network for crop yield prediction. This model is fitted on environmental data and management practices. The model was designed to capture the time dependencies of environmental factors and the genetic improvement of seeds over time without having their genotype information. The model can be extended with weather conditions, soil conditions, and management practices, which could explain the variation in crop yields. The proposed model achieves low error and outperforms all other methods that were tested. It was also shown that the model generalizes the yield prediction to untested environments without a significant drop in the prediction accuracy.

Many studies are devoted to the research and modeling of the international trade dynamics in crops. In [7], principal component analysis (PCA) was used to analyze the primary financial index data of some trade market companies, and the total score of the evaluation index was obtained in this study. Then, the financial and transaction indicators were simultaneously used as the input variables of the stock price prediction research.

The experiment results show that the neural network with the Bayesian regularization algorithm has the highest accuracy and can avoid overfitting in the training process. Finally, they compared the proposed PCA-BP neural network and an investment stock selection strategy based on the traditional stock selection analysis method and showed the proposed model is more effective. Article [8] compares the applicability of different models (autoregressive, regression, Holt-Winters) to predict market demand for dairy products based on historical data. Due to the small volume of training samples, the best results in terms of prediction accuracy were obtained using linear regression models.

In forecasting the dynamics of prices in agricultural markets, the most promising results were obtained using multilayer feedforward networks. The study [9] was devoted to assessing and comparing the applicability of ARIMA autoregressive models and multilayer neural networks for medium-term forecasting of the export level of cassava starch. Historical data on exports from Thailand from 2001 to 2003 were used as a pilot dataset. It was revealed that multilayer neural networks more accurately allow predicting the level of exports with higher accuracy. A network with a similar architecture was also used [10] for short-term forecasting of prices in the corn market. Another promising approach to forecasting is the use of recurrent neural networks. In the article [11], researchers proposed a nonlinear autoregressive exogenous model NARX that used a multilayer neural network architecture for predicting trade flows. NARX is a time-delayed feedforward network, that is, a network with no feedback loop, which can reduce its predictive ability. The main advantage of this model in comparison with the compositions of models trained using machine learning is the ability to jointly simulate linear and nonlinear dependencies between the inputs and outputs of the network. Paper [12] compares the NARX model with various models of time series analysis in forecasting the harvest area, production, and yield of soybean in Brazil. The results show that the NARX model provides reliable results and outperforms other methods. Although NARX catches non-linear temporal dependencies and shows good accuracy in time series prediction [13], the quality heavily depends on the hyperparameters of the time-delay component. They should be accurately fine-tuned with grid search or similar techniques, which is often hardly feasible. In the case of insufficient tuning of the hyperparameters or small datasets, the prediction accuracy can be lower than for simpler models. For example, the study [14] showed that in some scenarios radial based function (RBF) SVM regressors outperform the NARX models.

Article [15] presented a model for forecasting wheat production based on the recurrent network with the Long-Short Term Memory (LSTM) architecture [16]; such models are widely used for the medium-term analysis of time series. A comparison of the proposed model with several analogs was given. The results confirm that the proposed model achieved better forecasting accuracy than other methods, such as ARIMA and feedforward neural networks. A similar approach was also used in the article [17] for medium-term (up to 4 years) price forecasting in the grain market. LSTM networks can also be combined with convolutional networks (CNN) that makes them able to process raw data such as satellite images or spatial data together with the temporal features. Article [18] presents a CNN-LSTM model to predict soybean yield. The experiment shows that the model outperforms separate CNN and LSTM models. However, in the case of long time series, recurrent networks suffer from the gradient vanishing problem, which reduces their performance in comparison with the NARX models [19].

Transformer networks also seem to be prospective to forecast crop market dynamics and production level since they utilize a multi-head attention mechanism and catch long-term dependencies between various special and temporal features [20,21]. However, there are no studies of applying them for those problems yet.

Paper [22] integrates the analysis of trade flow indicators and the sentiment analysis of news articles. Obviously, raw text features cannot be passed directly to a prediction model. Thus they first build AffectiveSpace sentiment text embeddings [23] and use them together with the structured features to train a multilayer feedforward network. The experiments

show the proposed model is able to recognize the most meaningful changes in the market trend and to achieve positive returns during the simulation.

The review shows the mid-term wheat export, and production forecasting requires fairly complex models, which consider various features from the broad context of past time-steps and easily over-fit. Luckily, the studies such as [6] report that the overfitting can be significantly decreased with multiple regularization techniques, such as Bayesian regularization. The use of texts as the information sources for forecasting are also in the scope of research, but they can hardly be digested by prediction models and require applying information extraction techniques. Therefore, we will implement and test several context-aware models based on the sliding window approach, recurrent networks, and the Transformer architecture. Then we apply and test several regularization techniques to make those models more accurate.

3. Materials and Methods

3.1. Dataset

For simplicity we limit our study to the world's 15 top wheat exporters: Russia, Canada, the USA, France, Australia, Ukraine, Argentina, Romania, Germany, Kazakhstan, Bulgaria, Hungary, Czech Republic, Poland, and Lithuania. We also consider 15 primary wheat importers: Egypt, Indonesia, China, Turkey, Algeria, Bangladesh, Brazil, Philippines, Japan, Nigeria, Mexico, Morocco, Korea, Yemen, and Vietnam. We combine several data sources (Table 1):

- Wheat trade flows and production levels from UN FAOSTAT and UN Comtrade databases [24,25];
- Fuel prices. They affect the cost of using agricultural machinery; therefore affect the production level and prices;
- Macroeconomic indicators from International Monetary Fund [26] affect demand and consumption throughout the world;
- Normalized difference vegetation index (NDVI) from UN FAO [27] affect the production level. It usually helps to reveal the areas where vegetation is stressed. The values are obtained from the METOP-AVHRR sensor. The use of the whole raw data array would lead to unacceptable growth of the prediction models; therefore we consider per-seasonal county-level average and extreme (max, min) values.
- Topically-relevant messages from Twitter contain reports about deals in the international wheat market, climate anomalies, etc.

Figure 1 shows the approach we utilize to integrate all the heterogeneous features. The structured data can be normalized and passed directly to neural network-based regression models, while the social media data should be structured first.

Table 1. Feature set for the mid-term wheat export and production forecasting.

Feature Set	Frequency	For	Features
Trade flows	Annual	Country, Commodity	Export value
			Import value
			Re-export value
			Re-import value
Production	Annual	Country, Commodity	Production value
Fuel prices	Annual	World	WTI price, USD
			Brent price, USD

Table 1. Cont.

Feature Set	Frequency	For	Features
Macro-economic indicators	Annual	Country	Trade balance
			GDP
			Inflation (CPI)
			Inflation (PPI)
			Population
			Purchasing power parity (PPP)
NDVI indicators	Per season: max, min, median	Country	Normalized difference vegetation index
Sentiment	Annual	Country	Topic-wise summary tweet sentiment for a country

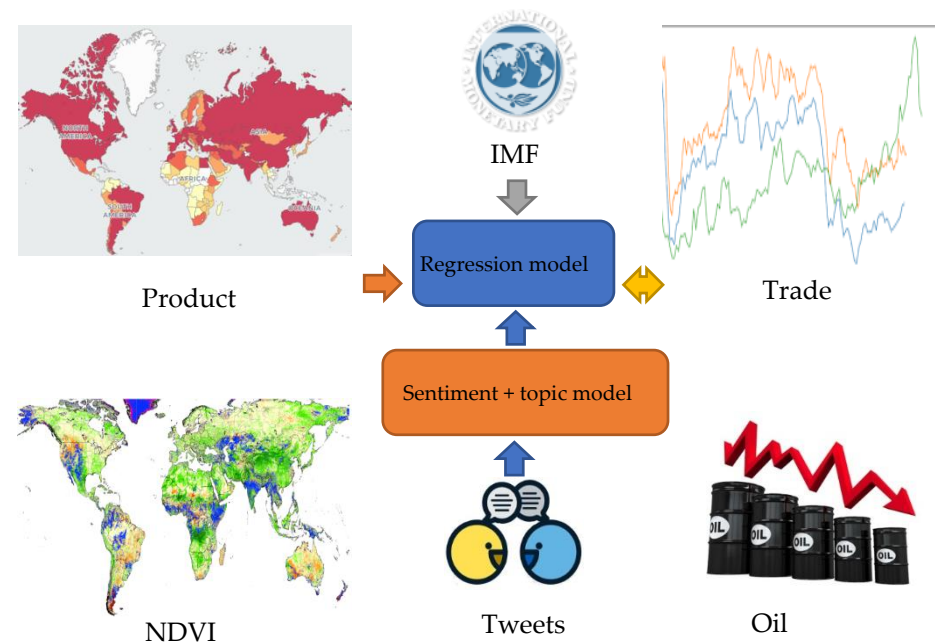


Figure 1. The data processing pipeline.

It would be naive to expect the tweets to contain precise country-level information about wheat production or trade flows, but they reflect some average market attitudes and expectations. Therefore, we evaluate the per-year country-level sentiment regarding wheat-related conversation topics and consider it as an additional feature to predict wheat production and trade flows. The final dataset is available on [28].

3.2. Modeling of Wheat Export

3.2.1. Regression Models

In our study we define the prediction of wheat export and production as a time-series analysis problem. Three modern neural network architectures were created and tested to solve that problem. The NARX-based model [10] is a feed-forward network (Figure 2a) with time-delay inputs, which has layers with *linear* and *elu* activations and predicts the future values for the export and production levels $Y_{E,P}(t, t+1, \dots)$ based on the inputs X and the output values $Y_{E,P}$ from the previous time-steps ($t-1, \dots, t-n$). In the experiments we have considered the last three-year history for each prediction; therefore, $n = 3$. The network should forecast the export Y_E and production Y_P two years ahead, and the iterative

prediction can lead to error accumulation. Therefore, we modified the model to return the outputs for the whole two-year period at once. During the experiments we utilized and evaluated different sets of features from the dataset to form the inputs X with the grid search approach. The inputs X for each time period contain the features of all the considered countries, because all the trade flows are strongly interrelated. We generate and concatenate simple ad-hoc country embeddings $X_{country}$ to the inputs to help the model distinguish feature groups related to different countries, then we process them with the standard Tensorflow Embedding layer to generate more compact real-valued embeddings. Trade indicators and production level in a particular country can affect the global wheat market significantly; therefore we use the self-attention mechanism to consider long-term dependencies between features. We also limited the model input by the 15 largest wheat exporters and importers to decrease the model complexity. The data processing algorithm in the network is the following. First, the model processes the input features with the dense network $d_{01}..d_{0k}$. The network has one layer with four neurons ($k = 4$) and *elu* activation. Then we calculate self-attention and concatenate it with the output of the $d_{01}..d_{0k}$ layer. After that we apply the global average pooling layer to form the attended average embedding of the input data (*Pooling*). Eventually, we process it with another multi-layer feed-forward network $d_1..d_m$ to form the outputs. The network has four layers ($m = 4$) with alternate *linear* and *elu* activations (except the last layer, which has *relu* activation to prevent the network from negative outputs) and four neurons in each layer.

As it is claimed above, the NARX networks require an accurate tuning of the sliding window (time delay) size, and, often it is not feasible because of hardware or time limitations. In addition, it is difficult to consider spatial and temporal dependencies with that architecture. The recurrent networks are free from those drawbacks.

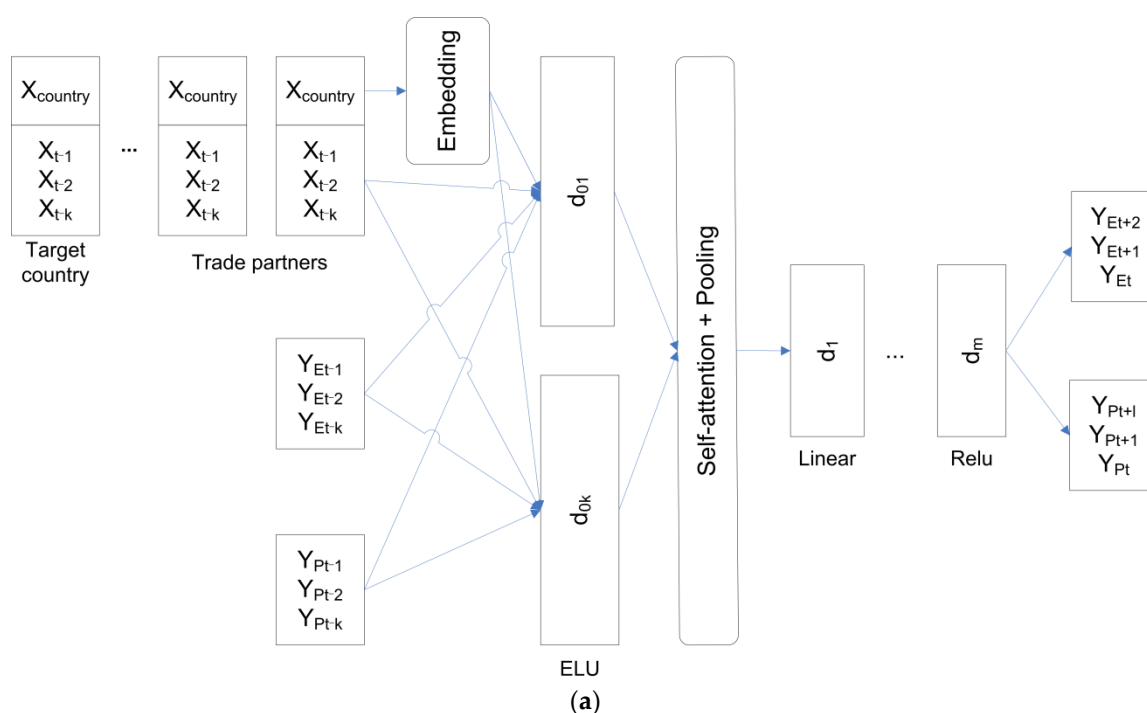


Figure 2. Cont.

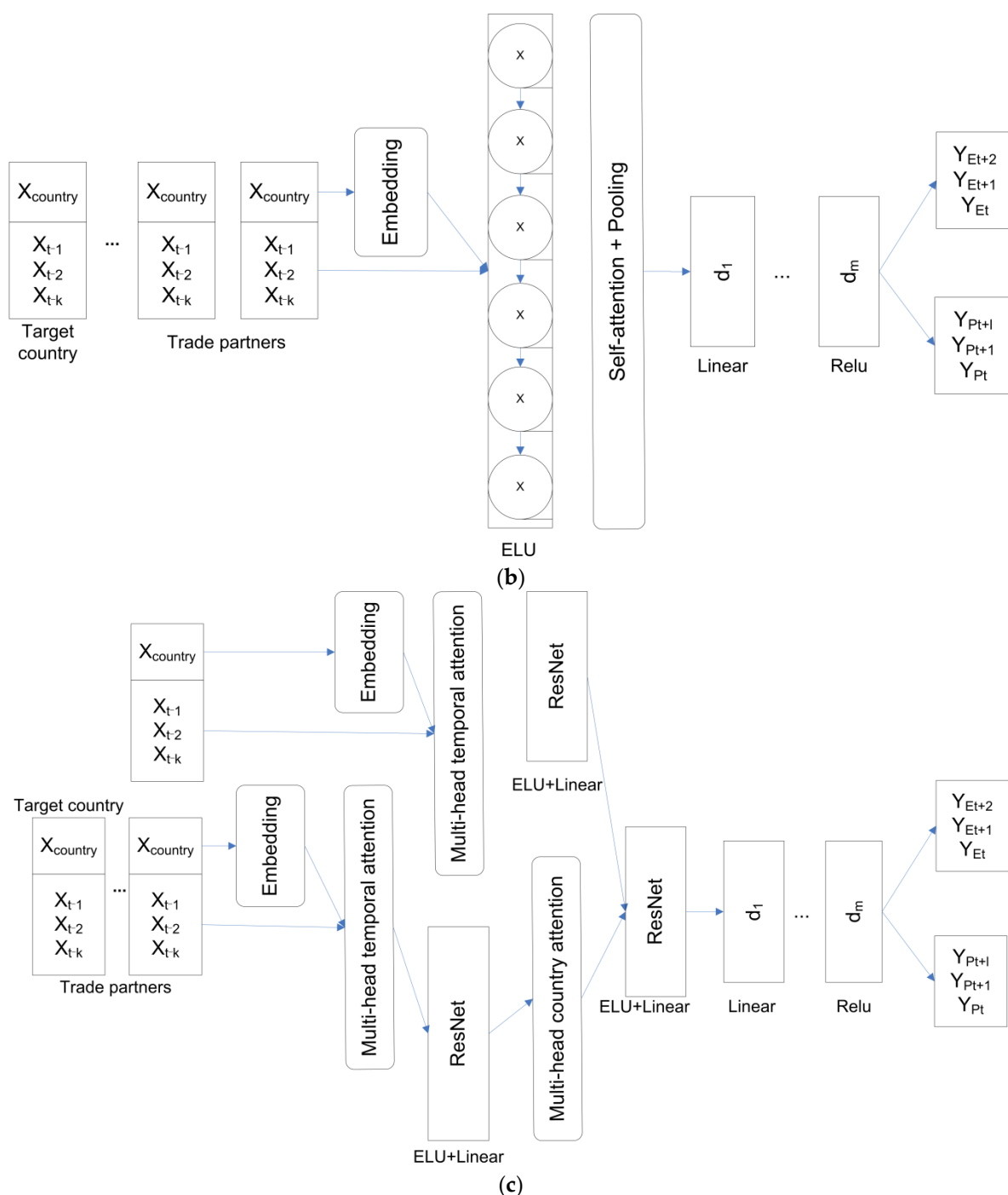


Figure 2. Neural network models to predict wheat export and production: (a) NARX model with self-attention. (b) recurrent network (LSTM/GRU/simple) with self-attention. (c) Transformer network with temporal and spatial attention.

In contrast to the NARX model, the recurrent models utilize all the past observations to make the predictions. The recurrent models have the same inputs, country embeddings and outputs as in the NARX model, except that they do not require past output values explicitly (Figure 2b). First, the inputs of the model are processed with the recurrent (LSTM, GRU, or simple recurrent [15]) network with *elu* activation and four neurons, which builds a single embedding for all the past observations. Then we transform that embedding with a multilayer feed-forward network $d_1 \dots d_m$. As in the NARX model, the network has four layers ($m = 4$) with alternate *linear*, *elu* and *relu* activations and four neurons in each layer. Similar to the NARX-based model, we apply the self-attention layer to consider

the dependencies between indicators of the countries. The network architecture after the recurrent layer is the same as for the NARX model.

The recurrent networks utilize backpropagation through time and have a gradient vanishing problem, affecting the prediction in a mid- and long-time series analysis. Instead of using backpropagation through time, the Transformer-like model (Figure 2c) applies positional encoding and multi-head attention mechanisms, which address the problem [20,29]. The Transformer network has the same inputs and outputs as the recurrent ones. First, the model generates embeddings for the inputs; they encode a country and a time step. We use the same embedding layer as for the previous models to encode a country, and apply a standard position encoding approach from [29] to generate the time step related information (positional embedding size is 3). Second, the model applies the multi-head attention to catch the dependencies between past time steps and passes the results through the residual network (ResNet). The all residual networks have four layers with alternate *linear* and *elu* activations, four neurons in each layers, and one residual connection between the input and the output of the ResNet. Eventually, the model uses multi-head attention to reveal dependencies between the series from different countries and generates the outputs with the network $d_1..d_m$ ($m = 4$). We use four-head self-attention for both countries and time steps.

We added l_2 regularization to all the layers of the models to make the models more stable. We also used a dropout mechanism to regularize the networks (dropout levels were chosen from the interval $[0..0.2]$). Adam optimizer with learning rate 0.001 was used to perform the models training. We also utilized the early stopping mechanism to prevent overfitting. We used the MSE loss function during the training because the prediction error was usually distributed normally. All the network hyper-parameters (number of network cells at each layer, number of layers, dropout level, regularization weights) were fine-tuned with the 3-fold cross-validation.

3.2.2. Stationary Term

Many approaches, such as ARIMA, represent the time series as a combination of stationary and non-stationary processes. We believe that would help generalize the series better, separating annual oscillations from long-term trends (Figure 3).

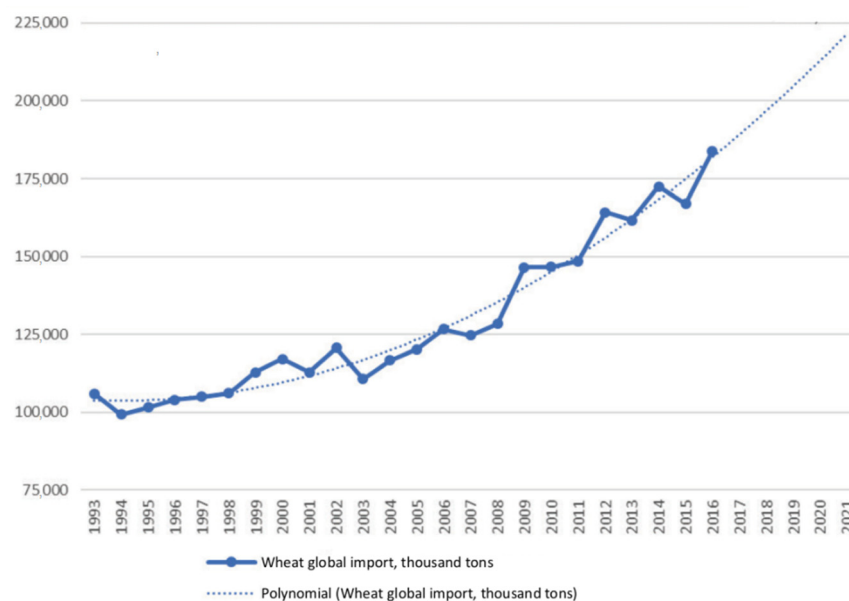


Figure 3. Annual global wheat import and its global trend (UN Comtrade data).

Therefore, we enhance our models with a stationary term (1).

$$Y_t = f(X_1, X_2, \dots, X_t) + \mathcal{E}(X_1, X_2, \dots, X_t) \quad (1)$$

where $X_1, X_2, \dots, X_t$ is a time series, $f(X_1, X_2, \dots, X_t)$ is a non-linear part (regression models a, b, c from the previous subsection), and $\mathcal{E}(X_1, X_2, \dots, X_t)$ is a stationary normally-distributed part. It is worth noting that we do not make any assumptions regarding the distribution of $f(X_1, X_2, \dots, X_t)$. We use a series of linear layers to estimate the E term. The idea is to train the model and find the layer weights $w_1, \dots, w_d$ to generate stationary weighted differences between various feature values in the time series (2).

$$\mathcal{E}(X_1, X_2, \dots, X_t) = w_d^T \left(\dots \left(w_1^T (X_1 X_2 \dots X_t) \right) \right) \quad (2)$$

The variance of $\mathcal{E}$ is unknown; therefore we add a tanh activation to keep the term in an interval $[-1..1]$ (3). That normalized term $\mathcal{E}_N$ can already be used to design an additive regularizer for the model loss. Eventually, we utilize Kullback–Leibler divergence $KL(\mathcal{P}(\mathcal{E}_N) || \mathcal{N}(0, 1))$ between the distribution of $\mathcal{E}_N$ in the time-series and normal distribution $\mathcal{N}(0, 1)$ as the regularizer (4).

$$\mathcal{E}_N = \tanh(\mathcal{E}) \quad (3)$$

$$l(X_1, X_2, \dots, X_t, Y_t) = \text{MSE}(Y_t, Y_t^*) + \alpha KL(\mathcal{P}(\mathcal{E}_N) || \mathcal{N}(0, 1)) \quad (4)$$

where the hyperparameter α is the weight of the regularization. This way, we force the models to generate the normally distributed term $\mathcal{E}$, which can be added to the smooth trend provided by $f(X_1, X_2, \dots, X_t)$. In the experiments we used fairly high weight for α , which is 10^5 to overpower the primary loss.

3.2.3. Sentiment Analysis and Topic Models

One of the prediction problems is a significant time gap between real-world events (large trade deals, economic crises, weather emergencies, etc.) and the appearance of information about them in global databases such as UN FAOSTAT. However, mass media and social networks report all those events almost immediately. Therefore, we try using that information to improve the prediction. We use Twitter for the experiments because it is available in all the top wheat exporters and importers and provides a search API. First, we collected all the tweets from 2012 to 2020 in English related to a “wheat grain” topic and the considered countries. Then we built a Latent Dirichlet Allocation (LDA) topic model to extract the primary topics of the tweets (Table 2). We used the Gensim library [30] and evaluated the perplexity and topic correlations to choose the topic number.

Table 2. Primary topics of the wheat grain related tweets (2012–2020).

Topic Name	Food Security	Wheat Pests	Yield/ Production Level	International Wheat Trade	Weather Conditions	Organic Agriculture
Keywords	area Asia bread curb prevent Europe farm feed India EU Australian	field crop bitly biggest impact countries problem	flour yields milling farming producers grains crops production harvest	export trade international importing harvesting prices imports bought buys tons market exports	farmers drought rain weather heavy	world GASC great organic breeding fertilizer

Eventually, we utilized the Polyglot library [31] to evaluate the sentiment of the messages. With Polyglot we define a function Sent that maps every token w from a token set W and every document d from a tweets collection D into an interval $[-1..1]$, where -1 stands for strong negative attitude, and $+1$ is for the solid positive attitude (5). We

have chosen that library because it is multilingual; therefore, our approach can be scaled to process texts on other languages.

$$Sent(w, d) : W \times D \rightarrow [-1..1] \quad (5)$$

Let D_{cy} be a set of tweets, related to particular country c and year y . We define a textual feature vector S , and each element of S is a weighted summary sentiment of the tweets related to a correspondent topic.

$$S(c, y) = (s_1, s_2, \dots, s_T), s_i = \sum_{d \in D_{cy}} \sum_{w \in d} p_{d,w} p_{w,i} \left(\sum_{s \in Ent(d)} Sent(s, d) \right) \quad (6)$$

where $Ent(d)$ is the set of all named entities (organizations) from tweet d , $p_{d,w}$ is the probability estimation that tweet d contains token w , and $p_{w,i}$ is the probability estimation that token w belongs to a topic i . They both are obtained from the topic model. Thus, for each country the vector direction shows sentiment of the tweets related to the topics, while the vector length reflects popularity of the topics.

The described methodology has the following limitations. We do not distinguish between tweets from the relevant institutions and messages from individuals. Additionally, we consider only tweets in English, which limits the completeness of the results. In the future, the official messages shall have a higher weight than those from individuals. We are also going to utilize a cross-lingual pipeline, which would help us to collect that information.

4. Results

First, we separated the dataset into the train and validation subsets. We used cross-validation on the train subset to set the hyper-parameters (dropout level, number of cells in the layers, weight of the regularization). Then we utilized bootstrapping to assess the models on the validation subset because most results are quite unstable. We applied a mean squared error MSE (7) and root relative squared error $RRSE$ (8) as the validation scores.

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - Y_i^*)^2 \quad (7)$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^N (Y_i - Y_i^*)^2}{\sum_{i=1}^N (\bar{Y} - Y_i^*)^2}} \quad (8)$$

Table 3 shows the two-year export prediction results. We included only the scores for the models with l2 layer regularization, because the models without regularization provide unstable results on bootstrap, which makes the any further comparison meaningless. The notation for the considered feature sets is the following.

Table 3. Export prediction scores for the baseline models.

Features	ALL		ALL—NDVIS		ALL—NDVIS—FUEL	
Models	$MSE, \times 10^{13}$	$RRSE$	$MSE, \times 10^{13}$	$RRSE$	$MSE, \times 10^{13}$	$RRSE$
GRU	0.63 ± 0.46	0.48 ± 0.41	1.85 ± 1.14	0.82 ± 0.65	1.33 ± 0.79	0.70 ± 0.54
LSTM	2.04 ± 1.57	0.87 ± 0.76	2.68 ± 1.47	0.99 ± 0.73	1.71 ± 1.39	0.79 ± 0.71
SimpleRNN	0.81 ± 0.59	0.54 ± 0.46	0.80 ± 0.61	0.54 ± 0.47	1.04 ± 0.81	0.62 ± 0.54
NARX	0.57 ± 0.35	0.45 ± 0.36	1.73 ± 0.56	0.80 ± 0.45	0.78 ± 0.56	0.53 ± 0.45
Transformer	0.62 ± 0.35	0.47 ± 0.36	1.18 ± 0.22	0.66 ± 0.28	0.79 ± 0.61	0.54 ± 0.47

“ALL”—trade flows, production levels, vegetation indexes, fuel prices, and macroeconomic indicators.

“ALL—NDVIS”—all the features above except vegetation indexes.

“ALL—NDVIS—FUEL”—all the features above except vegetation indexes and fuel prices.

It is worth noting that the best prediction scores were obtained for the models trained on the full feature set. That means both the vegetation indexes and fuel prices are crucial for the export prediction, although those overall are quite poor for real applications ($RRSE \approx 0.45$).

Table 4 presents the two year production level prediction results. All the results are fairly good, and the Transformer model shows the best prediction accuracy ($RRSE \approx 0.14$). We could not find any score improvement related to the NDVI indicators. We believe this is because the experimental dataset is quite small; therefore adding features leads to worse generalization. Additionally, there is a correlation between past production and vegetation levels, which makes the latter redundant. However, for most applications those vegetation indicators still would be crucial, because they are available almost immediately, while other indicators are collected by statistic agencies with a large gap.

Table 4. Production prediction scores for the baseline models.

Features	ALL		ALL—NDVIS		ALL—NDVIS—FUEL	
	<i>MSE</i> , $\times 10^9$	<i>RRSE</i>	<i>MSE</i> , $\times 10^9$	<i>RRSE</i>	<i>MSE</i> , $\times 10^9$	<i>RRSE</i>
GRU	1.80 ± 0.47	0.15 ± 0.07	1.88 ± 0.24	0.16 ± 0.05	1.68 ± 0.57	0.15 ± 0.09
LSTM	3.76 ± 0.41	0.22 ± 0.07	3.08 ± 0.36	0.20 ± 0.07	1.69 ± 0.37	0.15 ± 0.07
SimpleRNN	1.81 ± 0.30	0.15 ± 0.06	1.83 ± 0.27	0.15 ± 0.05	1.94 ± 0.53	0.16 ± 0.08
NARX	1.83 ± 0.29	0.15 ± 0.06	1.92 ± 0.57	0.16 ± 0.08	1.92 ± 0.36	0.16 ± 0.07
Transformer	1.77 ± 0.34	0.15 ± 0.06	1.56 ± 0.59	0.14 ± 0.08	1.82 ± 0.25	0.16 ± 0.06

Excluding macro-economic indicators has shown minor increase (≈ 0.02 – 0.03 for $RRSE$), both for export and production prediction errors.

Table 5 shows the two-year export prediction results for the models with stationary term and textual features. The notation for the considered feature sets is the following.

Table 5. Export prediction scores (regularization and sentiment features).

Features	ALL		ALL + SE		ALL + SE + SENTIMENT	
	<i>MSE</i> , $\times 10^{13}$	<i>RRSE</i>	<i>MSE</i> , $\times 10^{13}$	<i>RRSE</i>	<i>MSE</i> , $\times 10^{13}$	<i>RRSE</i>
GRU	0.63 ± 0.46	0.48 ± 0.41	0.89 ± 0.70	0.57 ± 0.50	0.23 ± 0.11	0.29 ± 0.20
LSTM	2.04 ± 1.57	0.87 ± 0.76	1.35 ± 0.81	0.70 ± 0.54	0.67 ± 0.38	0.49 ± 0.37
SimpleRNN	0.81 ± 0.59	0.54 ± 0.46	0.71 ± 0.15	0.51 ± 0.23	0.14 ± 0.03	0.22 ± 0.10
NARX	0.57 ± 0.35	0.45 ± 0.36	0.16 ± 0.05	0.24 ± 0.13	0.12 ± 0.03	0.21 ± 0.10
Transformer	0.62 ± 0.35	0.47 ± 0.36	0.14 ± 0.08	0.22 ± 0.17	0.13 ± 0.03	0.22 ± 0.10

“ALL + SE”—all the features, models with the stationary term.

“ALL + SE + SENTIMENT”—all the structural features plus textual ones, and models with the stationary term.

The table shows that the regularization helps to significantly improve the export prediction accuracy ($RRSE \approx 0.22$). In addition, it reduces the variance of the errors. Additional textual features also improve the predictions, especially for the NARX and Transformer models ($RRSE \approx 0.21$). In contrast to the export prediction models without regularization, those models are accurate enough to be utilized for information support of decision makers, who form state and large company agricultural policy.

Table 6 shows the two-year wheat production prediction results for the models with stationary term and textual features. The table shows the regularization helps to fairly improve the prediction accuracy of the wheat production. However, the textual features impair the prediction error. It seems the assessments of the market expectations and events are not worthy for the forecast of the wheat production. On the other hand, the production-related information appears in the social and text media as soon as in the statistical databases. In contrast to the export prediction, those features just increase the model's complexity, which leads to worse accuracy on the validation set.

Table 6. Production prediction scores (regularization and sentiment features).

Features	ALL		ALL + SE		ALL + SE + SENTIMENT	
Models	$MSE, \times 10^9$	RRSE	$MSE, \times 10^9$	RRSE	$MSE, \times 10^{10}$	RRSE
GRU	1.80 ± 0.47	0.15 ± 0.07	1.43 ± 0.49	0.13 ± 0.08	1.14 ± 0.10	0.39 ± 0.11
LSTM	3.76 ± 0.41	0.22 ± 0.07	1.69 ± 0.34	0.15 ± 0.07	1.26 ± 0.13	0.41 ± 0.13
SimpleRNN	1.81 ± 0.30	0.15 ± 0.06	1.87 ± 0.18	0.16 ± 0.04	1.02 ± 0.12	0.37 ± 0.13
NARX	1.83 ± 0.29	0.15 ± 0.06	1.93 ± 0.60	0.16 ± 0.09	1.23 ± 0.58	0.40 ± 0.27
Transformer	1.77 ± 0.34	0.15 ± 0.06	1.64 ± 0.42	0.14 ± 0.07	1.14 ± 0.12	0.39 ± 0.13

5. Discussion and Conclusions

This paper proposes several neural network models for mid-term forecasting of crop production and export and the approach to regularize them.

Our collected data set on the world's 15 largest exporters with heterogeneous features such as trade, macro-economic indicators of these countries, yields, and fuel prices could be as a basis for a comprehensive assessment of the global grain market, both in terms of export opportunities and more complete use of the resource potential for grain production.

The experiments show that the proposed NARX and Transformer models with regularization could be prospective methods for accurate prediction of crop production and export. It was also revealed that vegetation indexes and fuel prices are crucial for export prediction. Still, the fuel prices seem to be more important for predicting production than the NDVI indexes from past observations. We believe that good past NDVIs mean rich yield and stocks in those years, which could be used for export in the present or future. At the same time, the relationship between the past and future levels of NDVIs cannot be revealed with the models we tested. The experiments with the textual features show that they help increase trade flows prediction accuracy, but they could not improve the results of predicting production level. It seems that better results can be achieved when we consider different languages and vary messages depending on the importance of their sources.

The remaining errors of the export and production forecasting can be explained by ignoring some essential features. We plan to consider logistical and infrastructure conditions, climate indicators, and technological features in future research. The latter is a challenging problem, especially for large countries, where conditions vary significantly from region to region. In that case, representations of those features should reflect that variance, but keep low dimensionality to prevent the prediction models from overfitting.

Author Contributions: Conceptualization and methodology, Y.O.; investigation and validation, D.D.; writing-review and editing, D.D. and Y.O. All authors have read and agreed to the published version of the manuscript.

Funding: The study was financially supported by the Ministry of Education and Science of the Russian Federation (agreement no. 075-15-2020-805 from 2 October 2020).

Data Availability Statement: The data is available on http://keen.isa.ru/wheat_dataset (accessed on 29 September 2021).

Conflicts of Interest: The authors declare no conflict of interest the results.

References

1. Zurayk, R. Pandemic and food security. *J. Agric. Food Syst. Commun. Develop.* **2020**, *9*, 17–21.
2. Li, X.; Petropoulos, F.; Kang, Y. Improving forecasting with sub-seasonal time series patterns. *arXiv* **2021**, arXiv:2101.00827.
3. Sánchez-Durán, R.; Luque, J.; Barbancho, J. Long-term demand forecasting in a scenario of energy transition. *Energies* **2019**, *12*, 3095. [CrossRef]
4. Rundo, F.; Trenta, F.; di Stallo, A.L.; Battiato, S. Machine learning for quantitative finance applications: A survey. *Appl. Sci.* **2019**, *9*, 5574. [CrossRef]
5. Bhojani, S.H.; Bhatt, N. Wheat crop yield prediction using new activation functions in the neural network. *Neural Comput. Appl.* **2020**, *32*, 17. [CrossRef]
6. Khaki, S.; Wang, L.; Archontoulis, S.V. A cnn-rnn framework for crop yield prediction. *Front. Plant Sci.* **2020**, *10*, 1750. [CrossRef] [PubMed]
7. Jiasheng, C.; Wang, J. Exploration of stock index change prediction model based on the combination of principal component analysis and artificial neural network. *Soft Comput.* **2020**, *24*, 7851–7860.
8. Mor, R.; Bhardwaj, A. Demand forecasting of the short-lifecycle dairy products. In *Understanding the Role of Business Analytics*; Chahal, H., Jyoti, J., Wirtz, J., Eds.; Springer: Singapore, 2019; pp. 87–117.
9. Pannakkong, W.; Huynh, V.; Sriboonchitta, S. ARIMA versus artificial neural network for Thailand's cassava starch export forecasting. In *Causal Inference in Econometrics*; Springer: Cham, Switzerland, 2016; pp. 255–277.
10. Ayankoya, K.; Calitz, A.P.; Greyling, J.H. Real-time grain commodities price predictions in South Africa: A big data and neural networks approach. *Agrekon* **2016**, *55*, 483–508. [CrossRef]
11. Menezes, J.M.P., Jr.; Barreto, G.A. Long-term time series prediction with the NARX network: An empirical evaluation. *Neurocomputing* **2008**, *71*, 3335–3343. [CrossRef]
12. Abraham, E.R.; Mendes dos Reis, J.G.; Vendrametto, O.; Oliveira Costa Neto, P.L.D.; Carlo Toloi, R.; Souza, A.E.D.; Oliveira Morais, M.D. Time series prediction with artificial neural networks: An analysis using Brazilian soybean production. *Agriculture* **2020**, *10*, 475. [CrossRef]
13. Lee, W.K.; Tuan Resdi, T.A. Simultaneous hydrological prediction at multiple gauging stations using the NARX network for Kemaman catchment, Terengganu, Malaysia. *Hydrol. Sci. J.* **2016**, *61*, 2930–2945. [CrossRef]
14. Guzman, S.M.; Paz, J.O.; Tagert, M.L.M.; Mercer, A.E. Evaluation of seasonally classified inputs for the prediction of daily groundwater levels: NARX networks vs support vector machines. *Environ. Modeling Assess.* **2019**, *24*, 223–234. [CrossRef]
15. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.U.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM neural network based forecasting model for wheat production in Pakistan. *Agronomy* **2019**, *9*, 72. [CrossRef]
16. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
17. Liu, N.; Yu, J. Raw grain price forecasting with regression analysis. In *Proceedings of the 2019 International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)*, Wuhan, China, 23 June 2019; Atlantis Press: Paris, France, 2019.
18. Sun, J.; Di, L.; Sun, Z.; Shen, Y.; Lai, Z. County-level soybean yield prediction using deep CNN-LSTM model. *Sensors* **2019**, *19*, 4363. [CrossRef] [PubMed]
19. Guzman, S.M.; Paz, J.O.; Tagert, M.L.M. The use of NARX neural networks to forecast daily groundwater levels. *Water Resour. Manag.* **2017**, *31*, 1591–1603. [CrossRef]
20. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, Y. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *arXiv* **2019**, arXiv:1907.00235.
21. Chen, K.; Chen, G.; Xu, D.; Zhang, L.; Huang, Y.; Knoll, A. NAST: Non-Autoregressive Spatial-Temporal Transformer for Time Series Forecasting. *arXiv* **2021**, arXiv:2102.05624.
22. Picasso, A.; Merello, S.; Ma, Y.; Oneto, L.; Cambria, E. Technical analysis and sentiment embeddings for market trend prediction. *Expert Syst. Appl.* **2019**, *135*, 60–70. [CrossRef]
23. Cambria, E.; Hussain, A.; Havasi, C.; Eckl, C. Affectivespace: Blending common sense and affective knowledge to perform emotive reasoning. In *Proceedings of the WOMSA CAEPIA*, Seville, Spain, 9–13 November 2009; pp. 32–41.
24. Food and Agriculture Organization of the United Nations. Available online: <http://www.fao.org/faostat/en/> (accessed on 29 September 2021).
25. UN Comtrade: International Trade Statistics. Available online: <https://comtrade.un.org/data/> (accessed on 29 September 2021).
26. International Monetary Foundation. Available online: <http://www.imf.org/en/Data> (accessed on 29 September 2021).
27. Food and Agriculture Organization of the United Nations. Earth Observation. Available online: <http://www.fao.org/giews/earthobservation/country/index.jsp?code=AFG&lang=en> (accessed on 29 September 2021).
28. Dataset for Wheat Export and Production Forecasting. Available online: http://keen.isa.ru/wheat_dataset (accessed on 29 October 2021).
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
30. Rehurek, R.; Sojka, P. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valetta, Malta, 22 May 2010; pp. 46–50.
31. Al-Rfou, R. Polyglot: A Massive Multilingual Natural Language Processing Pipeline. Ph.D. Thesis, State University of New York at Stony Brook, New York, NY, USA, 2015.