water

MDPI

Article

A Deep Learning Model of Spatial Distance and Named Entity Recognition (SD-NER) for Flood Mark Text Classification

Robert Szczepanek

Institute of Geological Sciences, Faculty of Geography and Geology, Jagiellonian University, 30-387 Krakow, Poland; robert.szczepanek@uj.edu.pl

Abstract: Information on historical flood levels can be communicated verbally, in documents, or in the form of flood marks. The latter are the most useful from the point of view of public awareness building and mathematical modeling of floods. Information about flood marks can be found in documents, but nowadays, they are starting to appear more often on the Internet. The only problem is finding them. The aim of the presented work is to create a new model for classifying Internet sources using advanced text analysis (including named entity recognition), deep neural networks, and spatial analysis. As a novelty in models of this type, it was proposed to use a matrix of minimum distances between toponyms (rivers and towns/villages) found in the text. The resulting distance matrix for Poland was published as open data. Each of the methods used is well known, but so far, no one has combined them into one ensemble machine learning model in such a way. The proposed SD-NER model achieved an F1 score of 0.920 for the binary classification task, improving the model without this spatial module by 17%. The proposed model can be successfully implemented after minor modifications for other classification tasks where spatial information about toponyms is important.

Keywords: machine learning; ensemble model; cultural heritage; flood memory; Poland; open data; convolutional neural networks; natural language processing; high-water flood marks

check for updates

Citation: Szczepanek, R. A Deep Learning Model of Spatial Distance and Named Entity Recognition (SD-NER) for Flood Mark Text Classification. *Water* **2023**, *15*, 1197. https://doi.org/10.3390/w15061197

Academic Editor: Jianjun Ni

Received: 8 February 2023
Revised: 12 March 2023
Accepted: 16 March 2023
Published: 20 March 2023

1. Introduction

Searching for information on the Internet is not a simple task. The results returned by popular search engines for ambiguous phrases often contain unexpected results. In January 2023, out of the first 10 search results for the phrase "flood mark", only 5 in Google Search and 8 in DuckDuckGo refer to flood issues. Search results depend on the search engine used, and most importantly, they change over time [1]. With dozens or even hundreds of thousands of websites returned by search engines, it is extremely difficult to determine which of them contain information of interest to us. The task that each user faces is to classify the search results into results that meet their expectations and the others. Most often we do it intuitively based on previous experiences. For example, we ignore results that are described as advertising and those that meet the search criteria, but are not of interest to us. Fortunately, modern machine learning algorithms based on natural language processing (NLP) make this classification much easier [2–4], no matter what information we are looking for.

Flood marks (or high-water marks) are permanent graphic information describing and often also showing how high water reached during catastrophic floods [5,6]. Sometimes they are the only evidence of extreme events because no written records have survived. Quantitative information about flood events, especially those from many years ago, is often forgotten. This applies to both printed records and information about flood marks in the form of plaques usually mounted on the walls. This quantitative information is used in hydrological modeling, for example when determining potential flood hazard zones [7,8] or historic flood reconstruction [9]. Their reliability has been demonstrated, among others, in studies conducted by Bösmeier et al. [10]. It is also important that flood marks help

to build flood awareness in local communities, as is the case in the UK [11], Poland [12], Argentina, France, and New Zealand [13]. A few centuries ago, flood plaques were usually placed on the facades of churches, or in their interiors (Figure 1). Nowadays, they are placed on engineering structures such as bridges and viaducts, but also on private houses. For this reason, sometimes finding them is not easy. Few people know about the existence of even those publicly available. Information about some of them has not been described anywhere, and reaching them is a real challenge. Searching for information about flood marks in a small area, especially if it concerns historical marks, is feasible. An example may be the inventory of flood marks combined with the survey carried out for the city of Cracow (Poland) by Szczepanek et al. [14]. Some flood marks are created and installed in the form of standardized plaques by the meteorological services of many countries. Unfortunately, there is no central register of flood marks in Poland, and information about such marks is scattered over many places. Very often, such marks are created on the initiative of local communities or even individuals. Information about marks sometimes appears in the local press but can also be found on the Internet. Contemporary flood marks resulting from local initiatives are then original and unique, but finding them on the web is not easy.

Figure 1. Sample flood mark from 1813 attached to the side wall of the altar from Church of Divine Mercy at Felicjanek Street in Cracow. Inscription in Polish: "There was a great flood on 26 August 1813, to the point of the hand". (photo: R.Szczepanek, 2020).

In ancient times, information on flood marks was passed orally or sporadically appeared in printed materials [5]. Currently, the main source of information about the world is the Internet and social media. Searching for the right information should therefore be easier than before, but as practice shows, finding valuable and true information is not easy. Search engine results often contain pages unrelated to the content being searched for. This is due to many reasons, but one of them is the ambiguity of the searched phrases. Manually browsing the search results is time consuming, and adding more complex search terms can be cumbersome. For example, a search for "high water mark" (another name for flood mark) returns many economic links, not those related to flooding in rivers. This results directly from the ambiguity of the search term. In economics, a high water mark is the highest peak in value that an investment fund or account has reached. Whether a given text is related to hydrology or economics makes it much easier to decide on the basis of the context in which the expression appears. The search for the right context in the text has made tremendous progress in recent years through the use of machine learning and text analytics [4]. We decided to use these modern tools to search for information about floods on the Internet.

Searching and browsing websites manually is time consuming and very inefficient. The automation of the search process is therefore necessary, but even after that, the evaluation of search results' relevance remains the problem. The dynamic nature of information sources, such as blogs or discussion forums, makes the task even more difficult. The rapid development of machine learning (ML) in recent years has brought spectacular achieve-

ments in the field of computer vision and text analysis [15]. New ML algorithms have also proven their effectiveness in many hydrological tasks, in particular flood modeling [16]. However, it is not often in the earth sciences, which are dominated by quantitative analyzes (e.g., time-series forecasting [17]), that natural language processing (NLP) algorithms are used [18]. Typically, these are social media text analysis tasks during natural disasters [19].

The main goal of this research is to develop a method for the automatic classification of websites for searching for flood marks using machine learning. The main contributions of this research are listed as follows:

(1) The proposal of a novel deep learning ensemble model (SD-NER), which is a combination of named entity recognition with spatial distance analysis of the entities found in the text. The created model, tested on unstructured text from flood mark websites, achieved an F1 score of 0.920 for a binary classification task.

(2) The matrix of distances between 205 rivers and 30,700 towns and villages in Poland was calculated and published in CSV format under open license.

(3) A public and up-to-date repository of flood marks in Poland based on a bibliography study, online search and field survey was created and published as an interactive map on a dedicated web portal http://openhydrology.org/maps/flood_mark/ (accessed on 7 February 2023).

By the year 2023, about 300 existing flood marks were found and verified in Poland.

2. Related Work

2.1. Web Scraping

Web scraping is a technique to extract information from websites automatically [20]. Usually, when scraping tools are used on a website, it is assumed that the structure of the pages is known and repetitive. After determining the structure of the document, the necessary information is extracted from it in an orderly manner. With the help of web page parsers, such as lxml or BeatifulSoup for Python, a page with a known structure can be parsed to extract content from it [21]. However, when analyzing pages that are search results, knowledge of their structure cannot be assumed. Such a task becomes much more difficult than parsing pages with a known structure [22]. Due to the huge variety of website structures and tags used to describe them, the process of extracting the text itself is quite a challenge. Obtaining untagged plain text, with no HTML, CSS or JavaScript elements, is the first step.

In the next step, text from its conventional form consisting of letters and words must be converted to numeric form. This can be done in a number of ways. One of the classic methods from the year 2013 is word2vec, in which each distinct word is represented as a list of numbers called a vector [23]. The text from the website is converted into numerical tokens and then analyzed using the next model in the workflow. Those models range from simple naive Bayes [24] to the most extensive, such as BERT [25], GPT-3 [26], or recent ultra-large MT-NLG with 530 billion of parameters [27]. To describe the text in numerical form, multidimensional numerical representations based on large sets of text (corpora) are used. Increasingly, such generalized text representations are made available in the form of pre-trained models [28]. Machine learning (ML) models, and natural language processing (NLP) methods in particular, can support text analysis on many levels. In the context of website classification, two tasks seem to be the most useful: (a) named entity recognition tasks, which involve extracting words of a certain type from the text, for example, names of towns or rivers [29]; (b) classification tasks, which involve determining whether a given page is what we are looking for or not [30]. Website classification can be performed directly by task-oriented machine learning models [20,30] or using more generic models. Deep learning-based methods, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), outperformed traditional machine learning approaches, such as support-vector machine (SVM), in text classification tasks [31].

2.2. Toponym Extraction by Named Entity Recognition

Among the many types of machine learning models, NLP models are the most useful in the context of text analysis of web pages [32]. Named entity recognition (NER) is a task of identifying specific words or phrases ("entities") in the text and categorizing them, for example, as persons, locations, or events [29]. When looking for information about historical floods, we were interested in the fragments describing the relationships of inhabited areas (cities, villages) with the rivers whose waters overflowed. Such names that define places in geographical space (cities, rivers or mountain peaks) are called toponyms. The earliest works related to the use of NER were based on the use of simple geographic relationships between entities. In the work presented by Hu and Ge [33], a geographic knowledge base was modeled and constructed to support the toponym disambiguation procedure using a document collection consisting of 15,194 local Australian news articles. This experiment shows that the disambiguation accuracy is in the range of 74–85%, depending on the learning strategies used. The NER task concerning historical events is most often conducted on scanned documents, which are then passed to the digitization process [34,35]. The second common area of use for NER analysis is information about natural disasters posted by people on social media [31,36,37].

In 2021, Tempelmeier et al. [38] presented GeoVectors, a unique and comprehensive world-scale linked open corpus of OpenStreetMaps (OSM) entity embeddings covering the entire OSM dataset and providing latent representations of over 980 million geographic entities in 180 countries. This can be a very important step in the global use of toponyms as OSM is currently the richest publicly available information source on geographic entities worldwide. Geographic Question Answering (GeoQA) is an important and rapidly growing trend that connects Geographic Information Systems (GIS) with the NER [39]. Contractor et al. [40] proposed a joint spatio-textual reasoning model for answering tourism questions. Wang et al. [36] created NeuroTPR, a Neuro-net ToPonym Recognition model for extracting locations from social media messages. NeuroTPR extends a general recurrent neural network model for toponym recognition to address language irregularities in social media messages. Not only is the text analysis itself important, but its spatial and temporal context are also important [37]. This is particularly crucial in the case of natural disasters, when the timing of the events described plays a key role. Some of the NER models use the long short-term memory (LSTM) deep neural network architecture to detect toponyms [41]. An example of such a solution is the LSTM-CRF NER models proposed by Yadav et al. [42] or Dadas [43]. However, it should be noted that to the best of our knowledge, none of the authors used random websites as a source of analysis. In contrast to structured and rather short text-based sources (e.g., Twitter), websites are extremely diverse objects in terms of their structure.

Language corpora are the basis for text analyses. In the case of the Polish language, such a basis for analysis may be the NKJP (http://nkjp.pl/ (accessed on 7 February 2023)) [44,45]. The whole NKJP corpus consists of about 1.8 billion words. Entities in the NKJP are mostly manually annotated with metadata, i.e., they contain information about their origin, the title, the authors, etc. The second linguistic corpus to be used in the case of the Polish language may be the Grammar Dictionary of the Polish Language (SGJP; http://sgjp.pl/ (accessed on 7 February 2023)) [46]. Associated with this dictionary is the Morfeusz SGJP (http://morfeusz.sgjp.pl/ (accessed on 7 February 2023)) [47], a morphological analyser that enables NER analyses to be carried out on the text. Dadas [43] presented a neural architecture for NER in Polish and demonstrated how to improve its performance by using an entity-linking model with a knowledge base such as Wikipedia.

2.3. Spatial Distance

Spatial analyses on NER features are mainly focused on geocoding for features found in the text. Geocoding is the process of converting addresses or toponyms into geographic coordinates. Usually, finding a geographically referenced object (toponym) in text is an intermediate step. Most often, authors use geocoding to associate a found toponym with its

place in space [35]. A complete system for text geoparsing was proposed by Halterman [48] in the form of the Mordecai model built on word2vec [23] and Geonames. Kaczmarek et al. [49] presented an NER model with CNN and K-means clustering for spatial planning documents in Polish. As stated by the authors, the biggest struggle was the morphology of the Polish language. The concept of spatial nominal entity (SNoE) recognition was proposed by Medad et al. [50]. According to the authors, the same noun can be used to refer to a spatial object or other entities, which leads to ambiguities. Therefore, the recognition and disambiguation of spatial nominal entities in a corpus of an unstructured text compose a challenging problem. An interesting end-to-end ELECTRo-map method using statistical geocoding was proposed by Radford [51]. This model was evaluated on Wikipedia articles with coordinates linked with latitude and longitude. Examples of the processing and integration of spatial planning documents in Poland presented by Kaczmarek et al. [52] show that this is a dynamically developing area at the junction of the use of machine learning methods such as NLP and space management.

None of the found machine learning models based on NER used processed spatial information related to toponyms in their structure. The proposed model was tested for one country but combines the NER analysis with statistical measures of distance between rivers and towns.

3. Materials and Methods

3.1. Materials

All data used in this study are public, free, and in digital format. Two vector datasets are used as spatial data: (a) Hydrographic Map of Poland (MPHP50, 207 main rivers with names as lines) (Figure 2A); (b) State Register of Geographical Names (PRNG, 43,990 cities and villages as points) (Figure 2B).

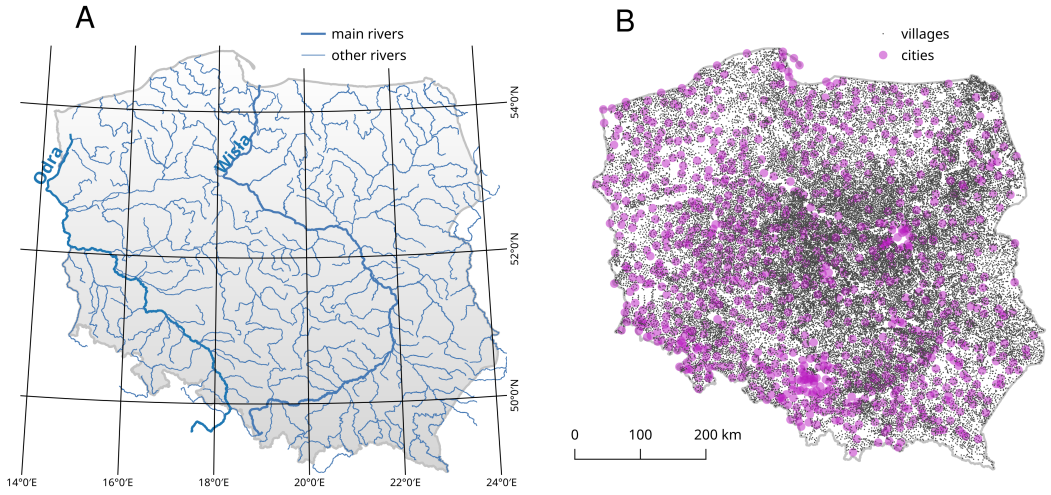

Figure 2. Spatial data for Poland used in this research. (**A**) Hydrographic Map of Poland (MPHP50) with hydrography network visualization of two main rivers: Wisła (Vistula) and Odra (Oder); (**B**) State Register of Geographical Names (PRNG) with location of cities and villages.

Google remains the most popular web search engine in 2023. In the proprietary tool category, it dominated the Internet many years ago. However, alternative tools such as Bing, Yahoo!, Ask, and Baidu are constantly being developed, and in some tasks, they can compete with the leader [1,53]. A separate group consists of tools that declare that they will provide the user with search privacy. These include DuckDuckGo and StartPage. Two popular search engines, one from each of these groups, were selected for the experiment: Google and DuckDuckGo.

We searched for the phrase "znak wielkiej wody", which means flood mark in Polish. The results in the form of URL addresses (links) were saved in text files. Data scraping was

carried out in January 2020, and in that time, Google reported about 8.3 million results. In January 2023, the number of results decreased to 6.3 million.

To create a balanced training set, the number of analyzed pages was limited to a few hundred initial results (Table 1). Almost four times as many links were fetched from Google than from DuckDuckGo because there were significantly more pages actually related to flood marks in the initial search results.

Table 1. Number of text sources used in this research.

Source Link	Scraped URLs	Scraped Documents (pdf)	Labeled as Flood Mark
Google	382	34	151
DuckDuckGo	110	6	30
Total	492	40	181

3.2. Methods

The starting points for the conducted analyses were information about floods found on websites. In addition to the classical text analysis, which will be described later, the proposed method uses toponyms extracted from the text. Toponyms refer to two types of objects: names of rivers and names of places. Both types of objects are shown in Figure 2.

The thesis put forward while creating and verifying the method is as follows: "If, in the text concerning the flood, we find the name of a river and the name of a town located nearby, a fluvial flood is probably described." Assuming that using NER tools we can extract toponyms from the text, it remains only to create a matrix describing the distances of all toponyms of these two types (rivers and towns) on a national scale. Information from such a matrix can be used in a machine learning model as an additional feature describing each text separately.

To create the distance matrix, source data on rivers and towns available in the form of Shapefile files were imported into the Postgres/PostGIS spatial database. Using the spatial function, `ST_Distance(river.wkb_geometry, town.wkb_geometry)` in this case, the shortest distance to each of the rivers in Poland was found for each town. An example of a distance calculation is shown in Figure 3.

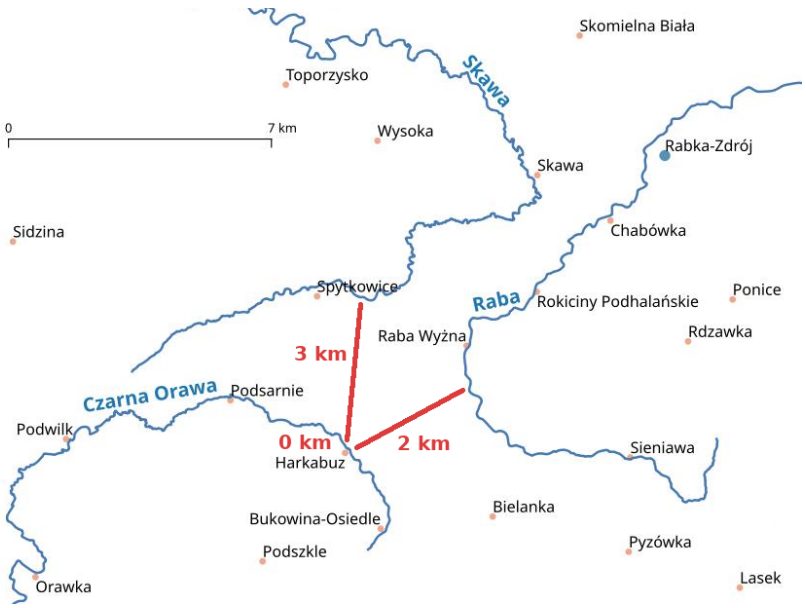

Figure 3. An example scheme for determining the distance. Minimum distance estimation between Harkabuz town and three visible rivers (Skawa, Raba, and Czarna Orawa). The town Harkabuz is on the Czarna Orawa river, so the distance is zero.

Dedicated Python scripts with the Beautiful Soup package were used for websites scraping, parsing, and tag removal. If the website link pointed to a pdf file, instead of Beautiful Soup, the pdfminer package was used to extract the raw text. Morfeusz, a morphological analyzer with pre-compliled Polish dictionaries, was used for lematization and tokenization [54]. Initial processing included the removal of stop words (based on https://github.com/stopwords-iso (accesed on 7 February 2023)) and lexemes shorter than 3 characters. The Morfeusz package was also used as the key named entity recognition (NER) tool for extracting toponyms from the text. In this particular case, the searched entities were defined as a 'geographical name' or 'part of a geographical name' ('nazwa_geograficzna' and 'człon_nazwy_geograficznej' in Polish). Only these elements contained the names of rivers, towns, and villages. After removing the duplicates, the names were saved for auxiliary spatial analysis. All URL links were manually tagged as containing or not containing flood mark information. Datasets were evenly balanced, with a similar number of links to pages with and without information about flood marks. A sample description of flood marks from a website is presented in Figure 4.

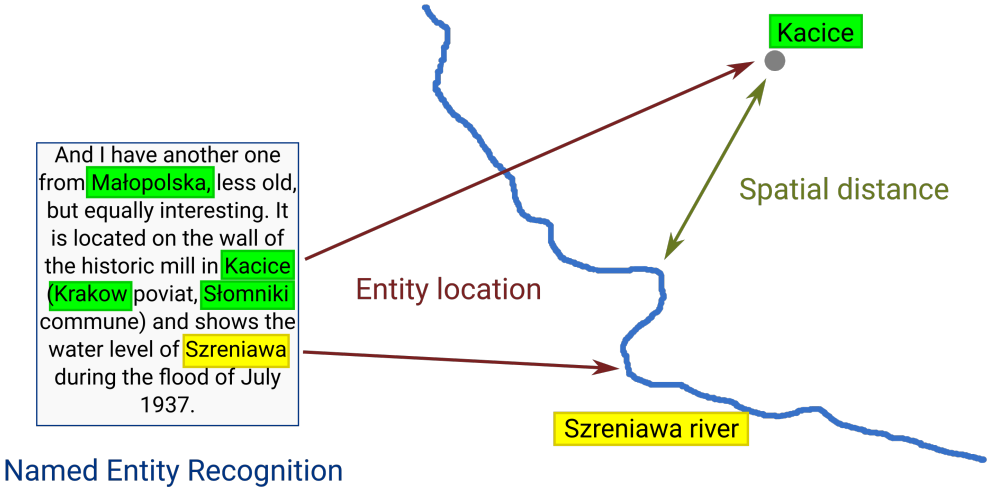

Figure 4. Sample description of flood mark (translation from Polish). Toponyms extracted from the text using NER are marked in green (towns) and yellow (river). Text source: "Time keepers" web portal—www.straznicyczasu.pl (accesed on 7 February 2023).

The last elements of the process of classifying websites provided by search engines were binary classifiers (Figure 5) determining whether a given source actually refers to the searched objects, in this case, flood marks. In this part of the workflow, three classifiers were used, the first two of which were used in a classic way by supplying information from the NLP model. The third classifier, in addition to the results from the second classifier, additionally used information about spatial distances between toponyms found in the text (Figure 5). All the elements of the workflow presented in Figure 5 are wrapped by dedicated scripts written in Python 3.6.

Three Python libraries were used to build machine learning models: scikit-learn, Tensorflow, and Keras [55]. The reference binary classification model was implemented using the naive Bayes MultinomialNB method from the scikit-learn library. This machine learning method implements the term frequency–inverse document frequency.

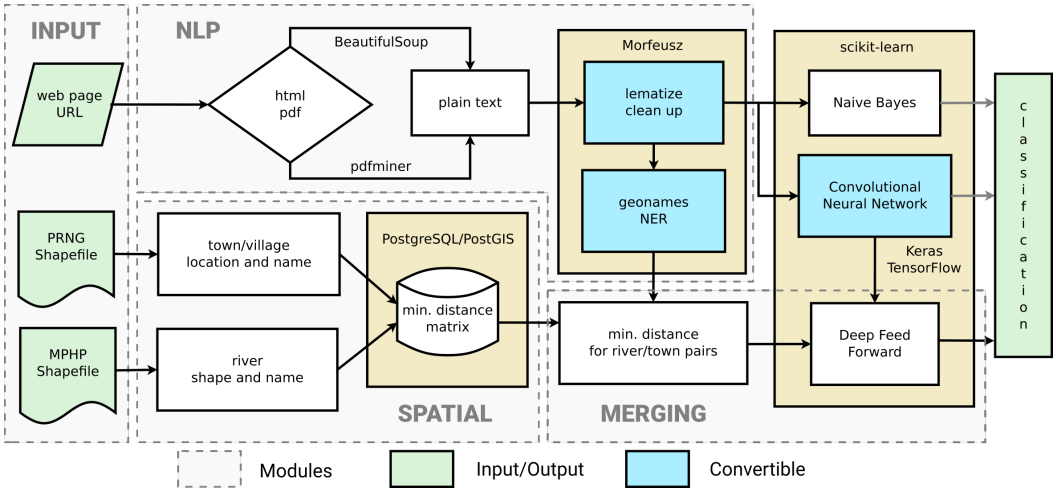

Figure 5. Proposed sequential SD-NER model workflow for binary web page classification based on toponym recognition and distance estimation.

The second model is a convolutional neural network (CNN) consisting of 7 sequential layers built in the Keras environment. The model has 1,661,352 trainable parameters and uses a batch size of 64. Rectified linear unit (ReLU) is used as the activation function, and binary cross-entropy is used with the Adam optimizer for the loss function.

The output from the CNN model is the final result of the binary classification but is also an input to consecutive deep feedforward (DFF) model (Figure 5). The main hypothesis is that information about rivers and residential areas (towns) affected by a flood can improve the classification if we can quantify the distance between the entities found in the text. Geographic names of rivers and towns (toponyms) are extracted by the named entity recognition module and merged with a distance matrix from the vector layers. The DFF sequential model has two features as input: the probability of a flood mark from the CNN and the normalized distance to the nearest river from the NER spatial analysis (Figure 5). The architecture of the DFF model consists of 4 sequential layers and has 42 parameters. This model uses ReLU and sigmoid activation functions with a batch size of 10 for calculations. Binary cross-entropy with the Adam optimizer is used for the loss function. We named this most complete path of the workflow the SD-NER model because it uses both the results of the text analysis (NER component) and spatial analysis (SD component). All three presented and tested models use machine learning in the final phase, and the last two use deep neural networks.

To maintain a large unprocessed dataset, the initial raw text dataset was equally divided (50/50) into training/verification and test datasets. In the training process, k-fold cross-validation was implemented with n_splits = 4 and shuffle = True to obtain better results and avoid overfitting the model.

To evaluate the results of binary classifiers, four metrics given in Equations (1)–(4) were used. The interpretation and designations of all metrics is shown in Figure 6.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

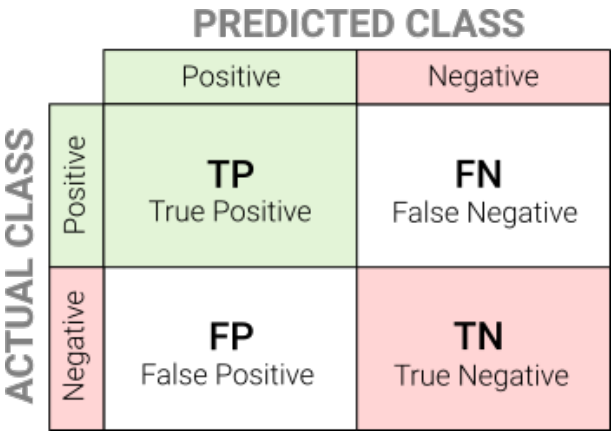

Figure 6. Confusion matrix for binary classification evaluation metrics.

Accuracy (1) is a metric for classification models that measures the number of predictions that are correct as a percentage of the total number of predictions that are made. Accuracy and loss are analysed to optimize the model structure and detect overfitting. This metric should not be used for unbalanced datasets, but that was not the case here. Precision (2) represents the number of correct positive results divided by the number of positive results predicted by the classifier. Recall (3) is the number of correct positive results divided by the number of all samples that should have been identified as positive. The F1 score (4) is the harmonic mean between precision and recall. It tells how many instances it classifies correctly, as well as if it does not miss a significant number of samples. The F1 score is probably the most sophisticated and comprehensive metric. High precision but lower recall gives extremely accurate output, but it then misses a large number of samples that are difficult to classify. The greater the given metric (in the range 0–1), the better the performance of the model is.

4. Results

As a result of the spatial analysis of vector data, a matrix of minimum distances between 205 rivers and 30,700 towns and villages in Poland was created. Distances have been rounded to full kilometers and are represented as integers for better performance. The data are saved in the CSV format (Figure 7) and have been released under an open license.

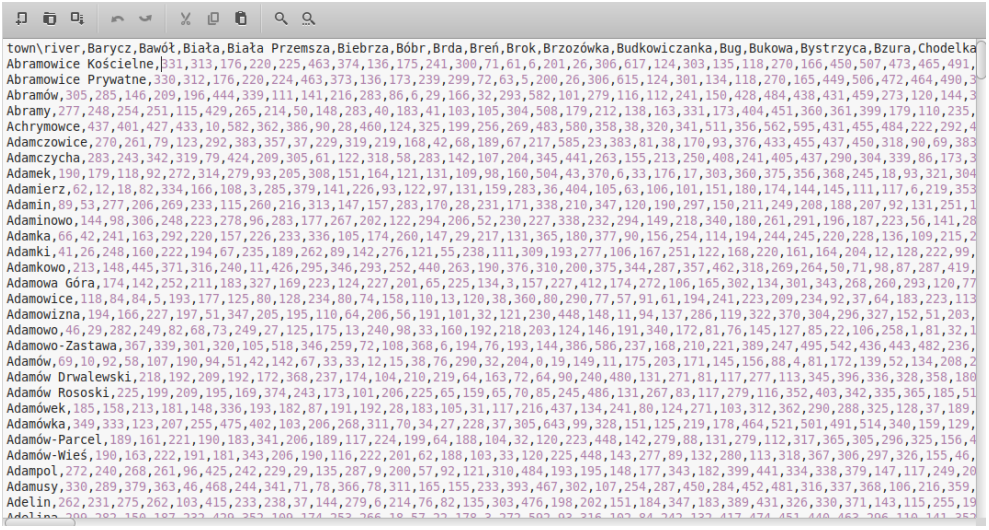
town\river,Barycz,Bawół,Biała,Biała Przemsza,Biebrza,Bóbr,Brda,Breń,Brok,Brzozówka,Budkowiczanka,Bug,Bukowa,Bystrzyca,Bzura,Chodelka
Abramowice Kościelne,331,313,176,220,225,463,374,136,175,241,300,71,61,6,201,26,306,617,124,303,135,118,270,166,450,507,473,465,491,
Abramowice Prywatne,330,312,176,220,224,463,373,136,173,239,299,72,63,5,200,26,306,615,124,301,134,118,270,165,449,506,472,464,490,3
Abramów,305,285,146,209,196,444,339,111,141,216,283,86,6,29,166,32,293,582,101,279,116,112,241,150,428,484,438,431,459,273,120,144,3
Abramy,277,248,254,251,115,429,265,214,50,148,283,40,183,41,103,105,304,508,179,212,138,163,331,173,404,451,360,361,399,179,110,235,
Achrymowce,437,401,427,433,10,582,362,386,90,28,460,124,325,199,256,269,483,580,358,38,320,341,511,356,562,595,431,455,484,222,292,4
Adamczowice,270,261,79,123,292,383,357,37,229,319,219,168,42,68,189,67,217,585,23,383,81,38,170,93,376,433,455,437,450,318,90,69,383
Adamczycha,283,243,342,319,79,424,209,305,61,122,318,58,283,142,107,204,345,441,263,155,213,250,408,241,405,437,290,304,339,86,173,3
Adamek,190,179,118,92,272,314,279,93,205,308,151,164,121,131,109,98,160,504,43,370,6,33,176,17,303,360,375,356,368,245,18,93,321,304
Adamierz,62,12,18,82,334,166,108,3,285,379,141,226,93,122,97,131,159,283,36,404,105,63,106,101,151,180,174,144,145,111,117,6,219,353
Adamin,89,53,277,206,269,233,115,260,216,313,147,157,283,170,28,231,171,338,210,347,120,190,297,150,211,249,208,188,207,92,131,251,1
Adaminowo,144,98,306,248,223,278,96,283,177,267,202,122,294,206,52,230,227,338,232,294,149,218,340,180,261,291,196,187,223,56,141,28
Adamka,66,42,241,163,292,220,157,226,233,336,105,174,260,147,29,217,131,365,180,377,90,156,254,114,194,244,245,220,228,136,109,215,2
Adamki,41,26,248,160,222,194,67,235,189,262,89,142,276,121,55,238,111,309,193,277,106,167,251,122,168,220,161,164,204,12,128,222,99,
Adamkowo,213,148,445,371,316,240,11,426,295,346,293,252,440,263,190,376,310,200,375,344,287,357,462,318,269,264,50,71,98,87,287,419,
Adamowa Góra,174,142,252,211,183,327,169,223,124,227,201,65,225,134,3,157,227,412,174,272,106,165,302,134,301,343,268,260,293,120,77
Adamowice,118,84,84,5,193,177,125,80,128,234,80,74,158,110,13,120,38,360,80,290,77,57,91,61,194,241,223,209,234,92,37,64,183,223,113
Adamowizna,194,166,227,197,51,347,205,195,110,64,206,56,191,101,32,121,230,448,148,11,94,137,286,119,322,370,304,296,327,152,51,203,
Adamowo,46,29,282,249,82,68,73,249,27,125,175,13,240,98,33,160,192,218,203,124,146,191,340,172,81,76,145,127,85,22,106,258,1,81,32,1
Adamowo-Zastawa,367,339,301,320,105,518,346,259,72,108,368,6,194,76,193,144,386,586,237,168,210,221,389,247,495,542,436,443,482,236,
Adamów,69,10,92,58,107,190,94,51,42,142,67,33,33,12,15,38,76,290,32,204,0,19,149,11,175,203,171,145,156,88,4,81,172,139,52,134,208,2
Adamów Drwalewski,218,192,209,192,172,368,237,174,104,210,219,64,163,72,64,90,240,480,131,271,81,117,277,113,345,396,336,328,358,180
Adamów Rososki,225,199,209,195,169,374,243,173,101,206,225,65,159,65,70,85,245,486,131,267,83,117,279,116,352,403,342,335,365,185,51
Adamówek,185,158,213,181,148,336,193,182,87,191,192,28,183,105,31,117,216,437,134,241,80,124,271,103,312,362,290,288,325,128,37,189,
Adamówka,349,333,123,207,255,475,402,103,206,268,311,70,34,27,228,37,305,643,99,328,151,125,219,178,464,521,501,491,514,340,159,129,
Adamów-Parcel,189,161,221,190,183,341,206,189,117,224,199,64,188,104,32,120,223,448,142,279,88,131,279,112,317,365,305,296,325,156,4
Adamów-Wieś,190,163,222,191,181,343,206,190,116,222,201,62,188,103,33,120,225,448,143,277,89,132,280,113,318,367,306,297,326,155,46,
Adampol,272,240,268,261,96,425,242,229,29,135,287,9,200,57,92,121,310,484,193,195,148,177,343,182,399,441,334,338,379,147,117,249,20
Adamusy,330,289,379,363,46,468,244,341,71,78,366,78,311,165,155,233,393,467,302,107,254,287,450,284,452,481,316,337,368,106,216,359,
Adelin,262,231,275,262,103,415,233,238,37,144,279,6,214,76,82,135,303,476,198,202,151,184,347,183,389,431,326,330,371,143,115,255,19

Figure 7. Head of matrix with minimum distance between rivers (columns) and towns (rows) in Poland. Published at https://doi.org/10.5281/zenodo.7618843 (accessed on 7 February 2023).

In the stage of web scraping, the main problem was the incorrect coding of websites (about 5% of all results). Although Polish phrases were searched, the results also included pages encoded in other languages. Text written in the Cyrillic, Chinese, and Arabic alphabets appeared on the analyzed pages. Manual decoding solved 95% of problematic text sources. In both search engines, the most important source turned out to be a portal dedicated to flood marks in Poland (Znaki Wielkich Wód), created by scientists from Toruń under the supervision of Marek Grześ (http://www.wielkawoda.umk.pl/ (accessed on 7 February 2023)). Information about flood marks was also found on discussion forums, private blogs, and static pages. Most of the flood marks found were located in the immediate vicinity of the two largest rivers in Poland, the Wisła and the Odra. The location and veracity of information about flood marks was verified by field research.

Despite the relatively small set of training data, stable results were obtained after about 200 epochs (Figure 8).

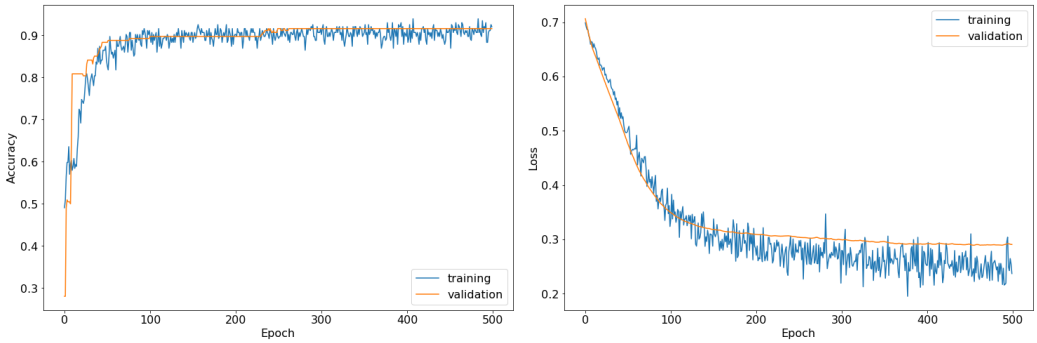

Figure 8. Accuracy and loss function during training and validation.

The results from the base model (naive Bayes) were significantly worse compared to the others, reaching an F1 score of 0.607. The second of the tested models (CNN), which used only information resulting from the text analysis, achieved an F1 of 0.786, which is 29% better than the base model. The last of the models, which in addition to the convolutional part (CNN) also used spatial information about the distance of objects, obtained an F1 of 0.920, i.e., 52% better than the base model and 17% better than the CNN model. The performances of all the models are presented in Table 2. With a very small amount of training data and a simple convolutional model, the training time of the models was negligibly short.

Table 2. Evaluation of tested models in binary text classification task.

Model	Accuracy	Precision	Recall	F1
Naive Bayes	0.829	0.850	0.472	0.607
CNN	0.868	0.810	0.772	0.786
SD-NER (proposed)	0.914	0.930	0.910	0.920

The obtained results confirm the assumption that the correct spatial interpretation of information about toponyms in the text can significantly improve the classification of the text. These differences are smaller for other metrics, except for recall. Recall improved from 0.472 all the way to 0.910 over the base model.

The vast majority of toponyms found in the text referred to objects only a few kilometers apart (Figure 9). Only in some cases was the distance greater than 40 km. This is consistent with the nature of floods in Poland, where fluvial floods are rather local and, despite significant flood losses, do not reach long distances.

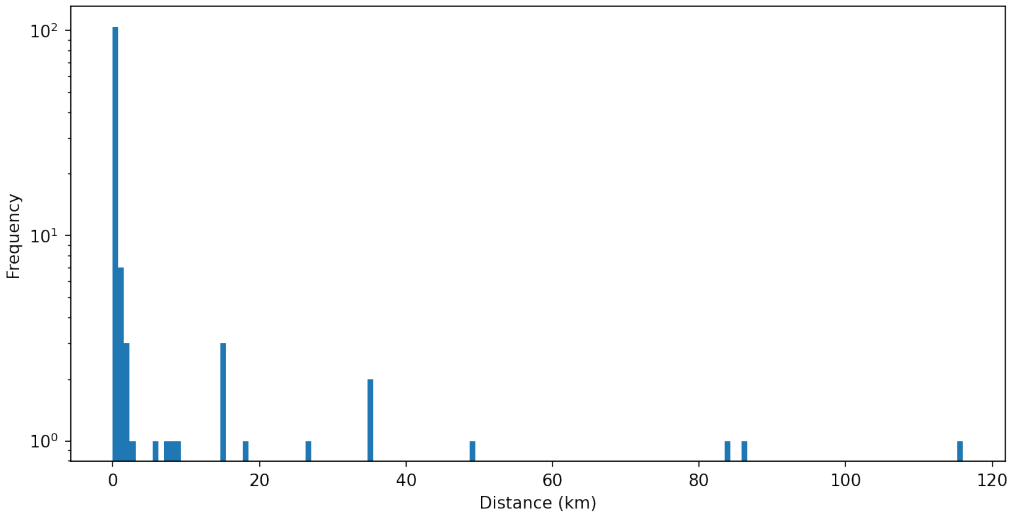

Figure 9. Histogram of distance between rivers and towns extracted from text. Note: The frequency (count) is on a logarithmic scale.

Probably not all distances shown in Figure 9 represent the cause and effect relationship (river-city) associated with the flood. Especially in the case of longer documents, the identified toponyms may have come from distant sentences. In all the documents examined, such unrelated correlations (for distances greater than 40 km) were detected in only four cases, including only one case over 100 km.

5. Discussion

The problem to which the presented analysis is devoted seemed to be practically impossible at the beginning. Based on a small sample of links to websites, we aimed to create a model that will correctly decide whether a given page actually concerns the sought issue, in our case, flood marks. The results were much better than expected, although quite simple machine learning methods were used. The biggest concern was related to the (assumed) lack of information about the structure of the website from which the information was collected. The variety of sources was huge, from discussion forums, blogs, and static pages to shared documents in pdf format. The use of popular methods of scraping and parsing available for the Python language enabled the initial preparation of the text for analysis by removing redundant tags and extracting the content of the documents. Certainly, by using information about page structures, much better results could be obtained, but it would not give the proposed method of assumed scalability. The results showed pages containing text in languages other than the base language (Polish). Cyrillic-, Chinese-, and Arabic-coded texts were encountered on the results pages. Sources that were indicated by search engines, but did not contain information about flood marks, most often concerned astrology, horoscopes, and religion, as well as music (names of albums or tracks). Some incorrect results should be associated with the specificity of the Polish language as well as heritage and cultural connotations. Toponyms were additionally extracted from the analyzed texts for further analysis. Similar methods have already been used by other authors who described the mechanism of geotagging tests based on text analysis. Panoutsopoulos et al. [56] described a model for the automatic identification and extraction of agricultural terms from unstructured text. However, the source of information was the AGRIS database, not the Internet. The authors used the spaCy library to analyze the texts. In the NER task, the model achieved an F1 of approximately 0.52 for agricultural terms. In the toponym geoparsing model described by Aldana-Bobadilla et al. [57], the authors used, i.a., the Corpus of Georeferenced Entities of Mexico (CEGEOMEX), yielding, for the global and local encoders, an F1 of 0.81. It should be noted that these tasks are

not identical to those presented in this article, so they cannot be directly compared with each other.

The second element of the proposed SD-NER ensemble model is a module related to deterministic spatial analyses based on vector layers for the area of Poland. Using the hydrography layer with major rivers and the settlement layer with towns and villages, the minimum distances for each pair of toponyms were calculated. Examples of distance histograms for three selected rivers in Poland are presented in Figure 10.

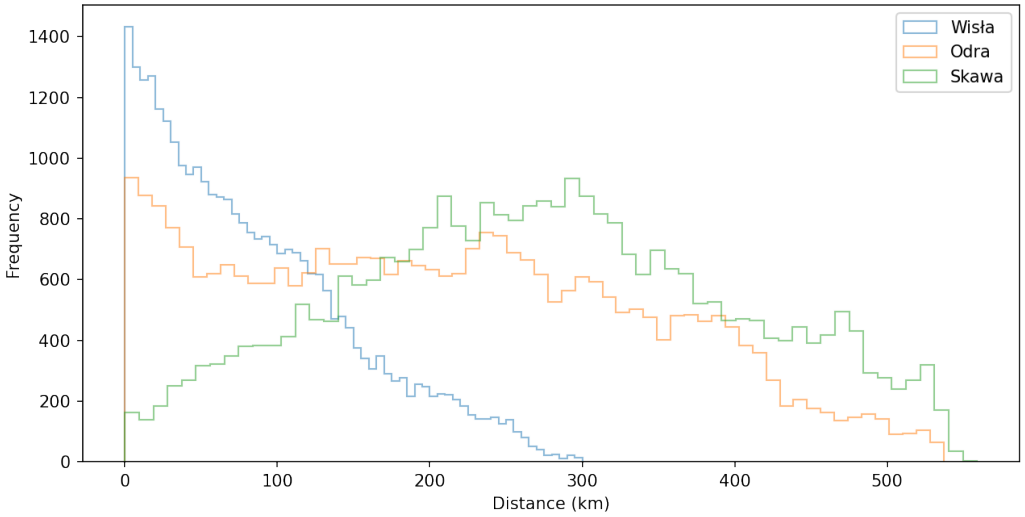

Figure 10. Histogram of distances from all towns in Poland to two main rivers (Wisła and Odra) and one smaller river from southern Poland (Skawa).

Although the analyses were carried out in the Postgres/PostGIS database environment, it was decided to publish the results in a simple CSV format under an open license so that other researchers could easily use them.

The Wisła (Vistula) river, which is the largest river in Poland, crosses practically the entire country, so the distances to towns do not exceed 300 km, and many towns are located on the river itself. A similar situation is in the case of the second largest river, the Odra (Oder), where many towns are located on the river itself. However, because the Odra river runs along the western border of the country, many towns are far away from it. Both rivers are the main source of the largest floods in Poland in the past, as well as in recent decades. The information about the distances between the toponyms is important because most of the floods are described in the texts as fluvial floods, i.e., connected with rivers. Therefore, if the geographical distance between the toponyms is large, the probability of cause and effect decreases significantly. Of course, this is not a rule, but it reflects the characteristics of the descriptions of historical floods quite well. The spatial dependence in the form of the distance between towns and villages in Poland and the two largest rivers is shown in Figure 11.

Poland's two main rivers, apart from a fragment in the south, flow through most of the country in a northerly direction and in a quite parallel fashion. The upper part of Figure 11 represents the settlements located east of the Wisła river, because the increase in the distance from this river is accompanied by the increase in the distance from the Odra river. The graph clearly shows an average distance between the two rivers of about 200–250 km. It is at such concentration points that sediments reach the X and Y axes on the graph. At the same time, these are areas with a distance close to zero, representing settlements lying directly on selected rivers. Settlements lying west of the Odra river are visible in the lower part of the graph on the right.

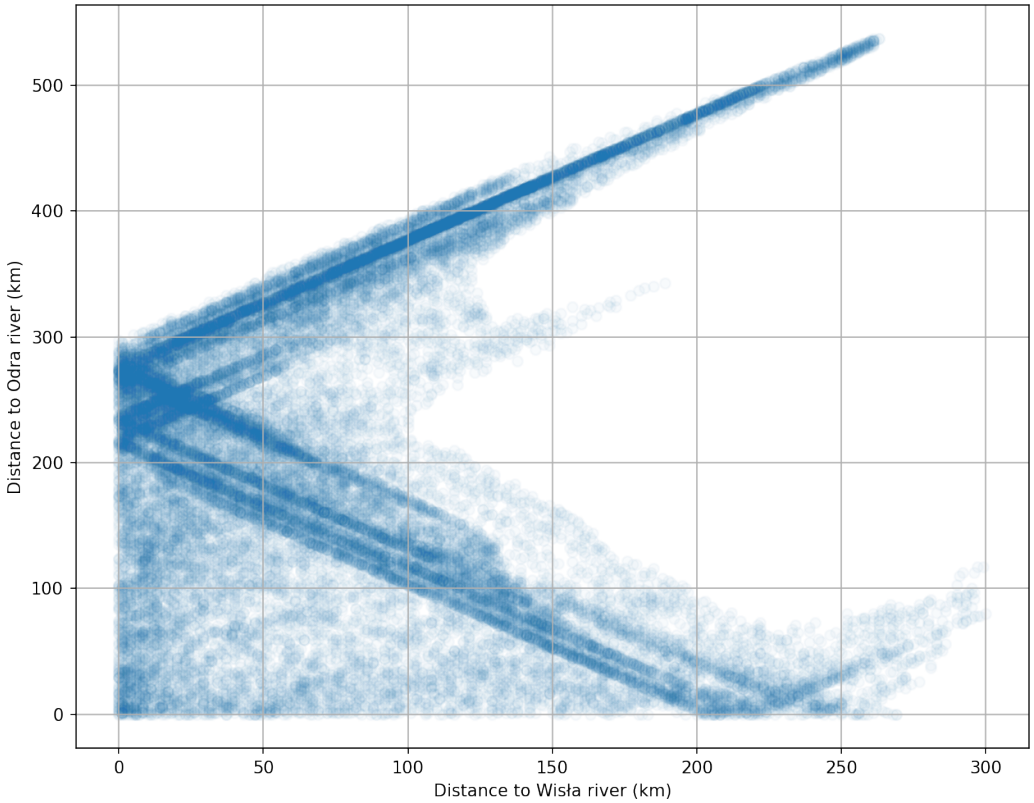

Figure 11. Scatter plot of distance from all towns/villages in Poland to two main rivers—Wisła and Odra. Each point represents one settlement. See Figure 2 for spatial references.

To the best of the authors' knowledge, analyses analogous to the one presented have not been published so far, and certainly not in relation to flood marks. In this sense, the proposed model is original and cannot be directly compared with the results of other authors. In the proposed model, it was possible to obtain a result comparable to more advanced machine learning models, such as BERT and its derivatives. The HerBERT and PolBERT model in the NER task currently achieve a score of 94.5 and 93.6, respectively [58,59]. They cannot be compared directly with the proposed model because NER is only one element of our model. Due to its specificity, Polish is more difficult to algorithmize than, for example, English. Local tools for lexical analysis dedicated to the Polish language are being developed, such as Morfeusz, which was used in this analysis, but the main trend is multilingual tools such as spaCy. Among the interesting works based on Polish and related to the presented issue, the works of Kaczmarek et al. [52] and Denisiuk et al. [60] should be mentioned.

The NLP is an emerging topic in water-related fields. In 2014, Murphy et al. [61] used NLP and NER on collections of newspaper articles from four cities in the U.S. Southwest to generate a network of water management institutions that reflect public perceptions of water management and the structure of water management in these areas. Faulkner et al. [62] ran a project based on NLP to identify topics and trends in the academic literature on the human right to water and sanitation. Tian et al. [63] conducted a study to investigate the use of NLP to handle customer complaints at the Water Utility Groningen in the Netherlands.

The proposed SD-NER method is universal because the searched phrase can refer to any topic (not only flood marks), and machine learning modules can be replaced in the future with better solutions. It is crucial to combine deterministic information about distances between objects with probabilistic information from NLP models. The final classification takes place in a model combining these two basic pieces of information. The deterministic component being the distance between toponyms significantly improves the obtained classification results. By adding a spatial module in the ensemble model, accuracy was improved by 0.440, precision was improved by 0.120, recall was improved by 0.138,

and the F1 score was improved by 0.134 to 0.920. Note that the CNN model was used as part of the proposed model. If more advanced models were used instead of a CNN, the improvement would probably be smaller.

The models of text analysis concerning floods and other threats presented in the literature are based on somewhat different assumptions. However, their results are comparable. Using a deep multi-branch BiGRU-CRF model, Fan et al. [64] achieved an F1 = 0.957 in the NER location (toponym) task for geological hazards. The toponym geocoding BiLSTM-CRF model proposed by Dewandaru [65], trained on about 645,000 news articles about, i.a., floods from Indonesian online news, obtained an average F1 score for the location of 0.892. The same class of model (BiLSTM-CRF) was used by Yuan [66] to extract spatio-temporal information from a Chinese archaeological site text. The authors obtained an F1 score of 0.879, but it should be noted that, as in the previous case, a very advanced machine learning model was used. The best results in toponym extraction, with an F1 of 0.961, were reported for the ALBERT model by Tao et al. [67] using NER in the Chinese language on the TPCNER dataset. The same authors reported F1 scores for the BiLSTM–CRF and BERT models of 0.860 and 0.921, respectively. These are currently among the most extensive models for this type of task. NLP was also used for the spatial analysis of flood problems on a global scale based on the available literature [68]. This can be considered the most generalized analysis. Perhaps the most practical project (THESPIAN-NER) in the field of text processing in the flood context was described by Bombini et al. [69]. The project implemented ArcheoNER and hsNER models that achieved F1 scores of 0.35 and 0.74, respectively. It can therefore be concluded that the SD-NER model proposed in this paper, thanks to the use of the distance between toponyms, achieves comparable and sometimes even better performance compared to more advanced models.

6. Conclusions

Information from the Internet, and in particular from social media is increasingly becoming an object of interest in the context of flood monitoring and prevention due to its scale and speed of transmission. Not all countries, such as the USA, maintain and update databases on flood marks [70]. Poland is not one of those countries. Valuable information about these marks can be found on the Internet, but with the vast amount of information contained therein, their proper categorization is difficult and time-consuming. The aim of the presented work was to test a new machine learning ensemble model that uses NLP methods and spatial analysis at the same time to classify text sources. The main challenge in the presented research was the extraction of text from a website with an unknown structure. Based on a search for flood marks carried out in Google and DuckDuckGo, it was shown that adding deterministic information about the minimum distance between toponyms to the SD-NER model can improve classification results by 17% for the F1 score. In the presented case of flood marks, these toponyms were the names of rivers and towns/villages. The architecture of the model is open, which makes it possible to replace the tested CNN model with more advanced models in the future. The presented task of the binary classification of Internet sources (pages, pdf files) is not limited to the presented issues. In an analogous way, the SD-NER model can be used for any other problem centered on toponyms. The resulting matrix of minimum river-town distances for the entire territory of Poland can be successfully used for other studies. One potential application could be to verify flood information posted on social media such as Twitter.

Funding: This research received no external funding.

Data Availability Statement: The created matrix of minimum distance between toponyms (rivers and towns) in Poland was published at https://doi.org/10.5281/zenodo.7618843 (accessed on 7 February 2023).

Acknowledgments: Thanks go to all open-source projects, such as Python, Postgres/PostGIS or QGIS, for providing the free tools that enabled this analysis.

Conflicts of Interest: The author declare no conflict of interest.

References

1. Dritsa, K.; Sotiropoulos, T.; Skarpetis, H.; Louridas, P. Search Engine Similarity Analysis: A Combined Content and Rankings Approach. In Proceedings of the International Conference on Web Information Systems Engineering, Amsterdam, The Netherlands, 20–24 October 2020; Springer: Cham, Switzerland, 2020; pp. 21–37. [CrossRef]
2. Jusoh, S. A study on NLP applications and ambiguity problems. *J. Theor. Appl. Inf. Technol.* **2018**, *96*, 6.
3. Dumbacher, B.; Diamond, L.K. SABLE: Tools for web crawling, web scraping, and text classification. In Proceedings of the Federal Committee on Statistical Methodology Research Conference, Washington, DC, USA, 7–9 March 2018.
4. Arnarsson, I.Ö.; Frost, O.; Gustavsson, E.; Stenholm, D.; Jirstrand, M.; Malmqvist, J. Supporting knowledge re-use with effective searches of related engineering documents—A comparison of search engine and natural language processing-based algorithms. In Proceedings of the Design Society: International Conference on Engineering Design, Delft, The Netherlands, 5–8 August 2019; Cambridge University Press: Cambridge, MA, USA, 2019; Volume 1, pp. 2597–2606. [CrossRef]
5. Pekárová, P.; Halmová, D.; Mitkova, V.B.; Miklánek, P.; Pekár, J.; Skoda, P. Historic flood marks and flood frequency analysis of the Danube River at Bratislava, Slovakia. *J. Hydrol. Hydromech.* **2013**, *61*, 326. [CrossRef]
6. Koenig, T.A.; Bruce, J.L.; O'Connor, J.; McGee, B.D.; Holmes R.R., Jr.; Hollins, R.; Forbes, B.T.; Kohn, M.S.; Schellekens, M.; Martin, Z.W.; et al. *Identifying and Preserving High-Water Mark Data*; Technical Report; US Geological Survey: Washington, DC, USA, 2016. [CrossRef]
7. Wyżga, B.; Radecki-Pawlik, A.; Galia, T.; Plesiński, K.; Škarpich, V.; Dušek, R. Use of high-water marks and effective discharge calculation to optimize the height of bank revetments in an incised river channel. *Geomorphology* **2020**, *356*, 107098. [CrossRef]
8. Grela, J. Assessment of the Potential Flood Hazard and Risk in the Event of Disasters of Hydrotechnical Facilities—The Exemplary Case of Cracow (Poland). *Water* **2023**, *15*, 403. [CrossRef]
9. Balasch, J.; Ruiz-Bellet, J.; Tuset, J.; Martín de Oliva, J. Reconstruction of the 1874 Santa Tecla's rainstorm in Western Catalonia (NE Spain) from flood marks and historical accounts. *Nat. Hazards Earth Syst. Sci.* **2010**, *10*, 2317–2325. [CrossRef]
10. Bösmeier, A.S.; Himmelsbach, I.; Seeger, S. Reliability of flood marks and practical relevance for flood hazard assessment in southwestern Germany. *Nat. Hazards Earth Syst. Sci.* **2022**, *22*, 2963–2979. [CrossRef]
11. McEwen, L.; Jones, O. Building local/lay flood knowledges into community flood resilience planning after the July 2007 floods, Gloucestershire, UK. *Hydrol. Res.* **2012**, *43*, 675–688. [CrossRef]
12. Gorączko, M. Flood Marks in Poland and Their Significance in Water Management and Flood Safety Education. In *Management of Water Resources in Poland*; Springer: Cham, Switzerland, 2021; pp. 253–267. [CrossRef]
13. Le Coz, J.; Patalano, A.; Collins, D.; Guillén, N.F.; García, C.M.; Smart, G.M.; Bind, J.; Chiaverini, A.; Le Boursicaud, R.; Dramais, G.; et al. Crowdsourced data for flood hydrology: Feedback from recent citizen science projects in Argentina, France and New Zealand. *J. Hydrol.* **2016**, *541*, 766–777. [CrossRef]
14. Szczepanek, R.; Toś, C.; Bodziony, M. Temporary flood marks proposal: What we learned after losing the baroque artifact from Cracow, Poland. *Int. J. Disaster Risk Reduct.* **2022**, *74*, 102942. [CrossRef]
15. Guo, J.; He, H.; He, T.; Lausen, L.; Li, M.; Lin, H.; Shi, X.; Wang, C.; Xie, J.; Zha, S.; et al. GluonCV and GluonNLP: Deep Learning in Computer Vision and Natural Language Processing. *J. Mach. Learn. Res.* **2020**, *21*, 1–7.
16. Sit, M.; Demiray, B.Z.; Xiang, Z.; Ewing, G.J.; Sermet, Y.; Demir, I. A comprehensive review of deep learning applications in hydrology and water resources. *Water Sci. Technol.* **2020**, *82*, 2635–2670. [CrossRef]
17. Szczepanek, R. Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost. *Hydrology* **2022**, *9*, 226. [CrossRef]
18. Maskey, M.; Ramachandran, R.; Miller, J.J.; Zhang, J.; Gurung, I. Earth science deep learning: Applications and lessons learned. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium; Valencia, Spain, 22–27 July 2018; IEEE: New York, NY, USA, 2018; pp. 1760–1763. [CrossRef]
19. Sit, M.A.; Koylu, C.; Demir, I. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: A case study of Hurricane Irma. *Int. J. Digit. Earth* **2019**, *12*, 11. [CrossRef]
20. Karthikeyan, T.; Sekaran, K.; Ranjith, D.; Vinoth, K.; Balajee, J. Personalized content extraction and text classification using effective web scraping techniques. *Int. J. Web Portals (IJWP)* **2019**, *11*, 41–52. [CrossRef]
21. Uzun, E.; Yerlikaya, T.; Kirat, O. Comparison of Python libraries used for Web data extraction. *Fundam. Sci. Appl.* **2018**, *24*, 87–92.
22. Plattner, T.; Orel, D.; Steiner, O. Flexible data scraping, multi-language indexing, entity extraction and taxonomies: Tadam, a Swiss tool to deal with huge amounts of unstructured data. In Proceedings of the Computation+ Journalism Symposium, Boston, MA, USA, 20–21 March 2016.
23. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.

24. Adiba, F.I.; Islam, T.; Kaiser, M.S.; Mahmud, M.; Rahman, M.A. Effect of corpora on classification of fake news using naive Bayes classifier. *Int. J. Autom. Artif. Intell. Mach. Learn.* **2020**, *1*, 80–92.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
26. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
27. Kharya, P.; Alvi, A. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model. 2021. Available online: https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ (accessed on 7 February 2023).
28. Yu, F.; Wang, D.; Shangguan, L.; Zhang, M.; Tang, X.; Liu, C.; Chen, X. A Survey of Large-Scale Deep Learning Serving System Optimization: Challenges and Opportunities. *arXiv* **2021**, arXiv:2111.14247.
29. Nadkarni, P.M.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 544–551. [CrossRef]
30. Kumar, S.A.; Nasralla, M.M.; García-Magariño, I.; Kumar, H. A machine-learning scraping tool for data fusion in the analysis of sentiments about pandemics for supporting business decisions with human-centric AI explanations. *PeerJ Comput. Sci.* **2021**, *7*, e713. [CrossRef] [PubMed]
31. Yu, M.; Huang, Q.; Qin, H.; Scheele, C.; Yang, C. Deep learning for real-time social media text classification for situation awareness—Using Hurricanes Sandy, Harvey, and Irma as case studies. *Int. J. Digit. Earth* **2019**, *12*, 1230–1247. [CrossRef]
32. Medlock, B.W. *Investigating Classification for Natural Language Processing Tasks*; Technical Report; University of Cambridge, Computer Laboratory: Cambdringe, UK, 2008. [CrossRef]
33. Hu, Y.H.; Ge, L. A Supervised Machine Learning Approach to Toponym Disambiguation. In *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*; Scharl, A., Tochtermann, K., Eds.; Springer: London, UK, 2007; pp. 117–128. [CrossRef]
34. Won, M.; Murrieta-Flores, P.; Martins, B. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Front. Digit. Humanit.* **2018**, *5*, 2. [CrossRef]
35. Viola, L.; Verheul, J. Machine Learning to Geographically Enrich Understudied Sources: A Conceptual Approach. In Proceedings of the ICAART (1), Valletta, Malta, 21–24 February 2020; pp. 469–475. [CrossRef]
36. Wang, J.; Hu, Y.; Joseph, K. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Trans. GIS* **2020**, *24*, 719–735. [CrossRef]
37. Scheele, C.; Yu, M.; Huang, Q. Geographic context-aware text mining: Enhance social media message classification for situational awareness by integrating spatial and temporal features. *Int. J. Digit. Earth* **2021**, *14*, 1–23. [CrossRef]
38. Tempelmeier, N.; Gottschalk, S.; Demidova, E. *GeoVectors: A Linked Open Corpus of OpenStreetMap Embeddings on World Scale*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 4604–4612. [CrossRef]
39. Mai, G.; Janowicz, K.; Zhu, R.; Cai, L.; Lao, N. Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions. *AGILE GISci. Ser.* **2021**, *2*, 1–21. [CrossRef]
40. Contractor, D.; Goel, S.; Singla, P. Joint Spatio-Textual Reasoning for Answering Tourism Questions. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 1978–1989. [CrossRef]
41. Plum, A.; Ranasinghe, T.; Orăsan, C. Toponym detection in the bio-medical domain: A hybrid approach with deep learning. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; pp. 912–921.
42. Yadav, V.; Laparra, E.; Wang, T.T.; Surdeanu, M.; Bethard, S. University of Arizona at semeval-2019 task 12: Deep-affix named entity recognition of geolocation entities. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; Association for Computational Linguistics: Cedarville, OH, USA, 2019; pp. 1319–1323. [CrossRef]
43. Dadas, S. Combining neural and knowledge-based approaches to named entity recognition in polish. In Proceedings of the Artificial Intelligence and Soft Computing, Zakopane, Poland, 16–20 June 2019; Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M., Eds.; Springer: Cham, Switzerland, 2019; pp. 39–50.
44. Przepiórkowski, A.; Bańko, M.; Górski, R.L.; Lewandowska-Tomaszczyk, B.; Łaziński, M.; Pęzik, P. National corpus of polish. In Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland, 21–23 April 2011; pp. 259–263.
45. Savary, A.; Piskorski, J. Language resources for named entity annotation in the National Corpus of Polish. *Control. Cybern.* **2011**, *40*, 361–391.
46. Woliński, M.; Saloni, Z.; Wołosz, R.; Gruszczyński, W.; Skowrońska, D.; Bronk, Z. *Słownik Gramatyczny Języka Polskiego*; SGJP: Warsaw, Poland, 2020.
47. Kieraś, W.; Woliński, M. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Pol.* **2017**, *XCVII*, 75–83.
48. Halterman, A. Mordecai: Full text geoparsing and event geocoding. *J. Open Source Softw.* **2017**, *2*, 91. [CrossRef]

49. Kaczmarek, I.; Iwaniak, A.; Świetlicka, A.; Piwowarczyk, M.; Harvey, F. Spatial Planning Text Information Processing with Use of Machine Learning Methods. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *6*, 95–102. [CrossRef]
50. Medad, A.; Gaio, M.; Moncla, L.; Mustière, S.; Le Nir, Y. Comparing supervised learning algorithms for spatial nominal entity recognition. *AGILE Gisci. Ser.* **2020**, *1*, 2020. [CrossRef]
51. Radford, B.J. Regressing Location on Text for Probabilistic Geocoding. *arXiv* **2021**, arXiv:2107.00080.
52. Kaczmarek, I.; Iwaniak, A.; Świetlicka, A.; Piwowarczyk, M.; Nadolny, A. A machine learning approach for integration of spatial development plans based on natural language processing. *Sustain. Cities Soc.* **2022**, *76*, 103479. [CrossRef]
53. Sheela, A.S.; Jayakumar, C. Comparative study of syntactic search engine and semantic search engine: A survey. In Proceedings of the 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 14–15 March 2019; IEEE: New York, NY, USA, 2019; Volume 1, pp. 1–4. [CrossRef]
54. Woliński, M. Morfeusz Reloaded. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 1106–1111.
55. Chollet, F. Keras, 2015. Available online: https://github.com/keras-team/keras (accessed on 7 February 2023).
56. Panoutsopoulos, H.; Brewster, C.; Espejo-Garcia, B. Developing a Model for the Automated Identification and Extraction of Agricultural Terms from Unstructured Text. *Chem. Proc.* **2022**, *10*, 94. [CrossRef]
57. Aldana-Bobadilla, E.; Molina-Villegas, A.; Lopez-Arevalo, I.; Reyes-Palacios, S.; Muñiz-Sanchez, V.; Arreola-Trapala, J. Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. *Remote Sens.* **2020**, *12*, 3041. [CrossRef]
58. Mroczkowski, R.; Rybak, P.; Wróblewska, A.; Gawlik, I. HerBERT: Efficiently pretrained transformer-based language model for Polish. *arXiv* **2021**, arXiv:2105.01735.
59. Kłeczek, D. Polbert: Attacking Polish NLP Tasks with Transformers. In Proceedings of the PolEval 2020 Workshop, Warsaw, Poland, 26 October 2020; pp. 79–88. Available online: http://poleval.pl/files/poleval2020.pdf (accessed on 7 February 2023)
60. Denisiuk, A.; Ganzha, M.; Wasielewska-Michniewska, K.; Paprzycki, M. Feature Extraction for Polish Language Named Entities Recognition in Intelligent Office Assistant. In Proceedings of the HICSS, Maui, HI, USA, 4–7 January 2022; pp. 1–10.
61. Murphy, J.T.; Ozik, J.; Collier, N.T.; Altaweel, M.; Lammers, R.B.; Kliskey, A.; Alessa, L.; Cason, D.; Williams, P. Water relationships in the US southwest: Characterizing water management networks using natural language processing. *Water* **2014**, *6*, 1601–1641. [CrossRef]
62. Faulkner, C.M.; Lambert, J.E.; Wilson, B.M.; Faulkner, M.S. The human right to water and sanitation: Using natural language processing to uncover patterns in academic publishing. *Water* **2021**, *13*, 3501. [CrossRef]
63. Tian, X.; Vertommen, I.; Tsiami, L.; van Thienen, P.; Paraskevopoulos, S. Automated Customer Complaint Processing for Water Utilities Based on Natural Language Processing—Case Study of a Dutch Water Utility. *Water* **2022**, *14*, 674. [CrossRef]
64. Fan, R.; Wang, L.; Yan, J.; Song, W.; Zhu, Y.; Chen, X. Deep learning-based named entity recognition and knowledge graph construction for geological hazards. *ISPRS Int. J. Geo Inf.* **2019**, *9*, 15. [CrossRef]
65. Dewandaru, A.; Widyantoro, D.H.; Akbar, S. Event geoparser with pseudo-location entity identification and numerical argument extraction implementation and evaluation in Indonesian news domain. *ISPRS Int. J. Geo Inf.* **2020**, *9*, 712. [CrossRef]
66. Yuan, W.; Yang, L.; Yang, Q.; Sheng, Y.; Wang, Z. Extracting Spatio-Temporal Information from Chinese Archaeological Site Text. *ISPRS Int. J. Geo Inf.* **2022**, *11*, 175. [CrossRef]
67. Tao, L.; Xie, Z.; Xu, D.; Ma, K.; Qiu, Q.; Pan, S.; Huang, B. Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS Int. J. Geo Inf.* **2022**, *11*, 598. [CrossRef]
68. Zhang, M.; Wang, J. Global Flood Disaster Research Graph Analysis Based on Literature Mining. *Appl. Sci.* **2022**, *12*, 3066. [CrossRef]
69. Bombini, A.; Alkhansa, A.; Cappelli, L.; Felicetti, A.; Giacomini, F.; Costantini, A. A Cloud-Native Web Application for Assisted Metadata Generation and Retrieval: THESPIAN-NER. *Appl. Sci.* **2022**, *12*, 12910. [CrossRef]
70. Ning, H.; Li, Z.; Hodgson, M.E.; Wang, C. Prototyping a social media flooding photo screening system based on deep learning. *ISPRS Int. J. Geo Inf.* **2020**, *9*, 104. [CrossRef]