TOWARDS AN EFFECTIVE APPROACH FOR CYANOBACTERIA AFFECTED LOCATIONS EXTRACTION FROM NEWS FEEDS

by

ANUJA CHANDRAKANT JADHAV

(Under the Direction of Lakshmish Ramaswamy)

ABSTRACT

Blue-Green Algae (BGA) are a toxic phytoplankton which are now a worldwide phenomena and a concern to various public authorities. The issue of cyanobacteria was brought to a highlight with its first ever official published report in 1878. The thick belts of these BGA (also known as Cyanobacteria) restricts the sunlight from penetrating into the water bodies which leads to depletion of the levels of dissolved oxygen thus hampering the aquatic life as well as affecting the humans coming in contact of the contaminated water. The year 2016 has seen extensive algal bloom for some prominent and large water bodies like Lake Okeechobee. With our work we aim to extract information of locations affected by cyanobacteria from news articles. We first illustrate the purpose of analyzing news articles with respect to BGA and, then formulate this purpose into an extraction workflow using unsupervised machine learning and named entity recognition framework based data mining approach. We evaluate our approach by validations from supervised learning techniques on data set gathered from news articles for the year 2017. Our experiment shows that our approach is effective for the locations detection problem statement in the cyanobacterial news articles domain.

INDEX WORDS: CyanoHAB, Blue Green Algae, Red Tide, Cyanobacteria, Machine Learning, Text Classification, Named Entity Recognition

TOWARDS AN EFFECTIVE APPROACH FOR CYANOBACTERIA AFFECTED LOCATIONS EXTRACTION FROM NEWS FEEDS

by

ANUJA CHANDRAKANT JADHAV

B.E., University of Mumbai, India, 2012

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2018

TOWARDS AN EFFECTIVE APPROACH FOR CYANOBACTERIA AFFECTED LOCATIONS EXTRACTION FROM NEWS FEEDS

by

ANUJA CHANDRAKANT JADHAV

Major Professor:	Lakshmish Ramaswamy
Committee:	Ismailcem Budak Arpinar
	Shannon Quinn

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

DEDICATION

To all the entities from across the world who strive for the protection, conservation and preservation of Earth's natural resources.

ACKNOWLEDGMENTS

I would sincerely like to thank my advisor Dr. Ramaswamy, for giving me an opportunity to be a part of Data Intensive and Pervasive Systems lab's one of the valuable projects, CyanoTracker Web Application. This opportunity was a great learning curve for me towards Machine Learning algorithms applications. The CyanoTracker Web application project is partially funded by National Science Foundation (NSF) under grant number CCF-1442672 and I would like to extend my sincere gratitude to NSF.

I would like to extend my heart-felt thanks to my committee members, Dr. Shannon Quinn for scheduling a meeting to discuss, provide helpful insights especially during a busy phase of the semester and Dr. Ismailcem Budak Arpinar for your suggestions on Semantic Triple Extraction techniques to enhance this work with a fresh approach.

Last but certainly not least, I would like to thank my family for their incredible love and support. I am grateful to all my friends but a special thanks to my friend and an extremely kind and caring flatmate, Ankita Joshi for all the support. Thank you so much for all the suggestions especially for the defense presentation slides. I would also like to thank my other friends whom I have troubled the most with my queries in my research journey. Dr. Vinay Kumar Boddula, a former PhD student of Dr. Ramaswamy for his feedback and insights, Narita Pandhe for giving me a heads up on the machine learning basics and patiently answering all my queries, Raunak Dey for proactively volunteering to proofread my thesis and providing suggestions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Cyanobacteria has emerged as a serious water quality issue in recent years. It was a concern since its first ever official reporting in 1878 [3] [4]. But with recent years the issue has been aggravating and hampering remarkable number of natural water resources. Many water bodies have been endangered and led to being closed for restoration [5]. The issue further has been creating reports of underwater habitat being endangered. Numerous locations have seen dead fishes and bird sickening behavior due to the cyanobacteria [6]. Harmful cyanobacteria adversely affect the wildlife, livestock and animals around the water bodies. Blooms that occur on a massive level or prominently known water areas get immediate attention. But there are a lot more water bodies which may be visited sparsely and hence would go unnoticed for its bloom issue.

With University of Georgia's initiative of creating an early warning system to track algal bloom, CyanoTracker web application [7] was developed for monitoring CyanoHABs in lakes and ponds across Georgia, USA. The application aims at integrating remotely sensed data with various social media platforms to capture more information. These CyanoHABs found on the surface of water bodies are visible to naked eyes and hence could be reported by anyone through online social media sites like Facebook, Instagram and Twitter. As of today the project has four sensors deployed at different lakes across Georgia, USA. The data collected from these sensors is integrated with social media data to capture more information. With dedicated hashtags like $CyanoHABs$, $Cyanobacteria$, $Microcystin$, $BlueGreenAlgae$ and $RedTide$ the application tracks various algal bloom posts thus taking advantage of social sensing.

For our research, we have utilized news articles. New articles provide enormous content on the issue of cyanobacteria. We came across many articles which not only mentioned about cyanobacteria reporting in different locations but also highlighted other related issues like fish, dogs and other wildlife death incidents, closing down of beach activity. New articles are moreover drafted and published by designated journalists which reduces the colloquial language usage unlike other social platforms like Twitter and Facebook. Validation of a news article is always possible since the details of source agency is available. Hence we can choose to consider an article only after validating a source.

1.1 Thesis Contribution

In our research study, we explore a set of research questions:

1. How can we extract content of news articles?
2. How can news articles be classified using various machine learning approaches?
3. How can location information be extracted from given news article?
4. What location modelling approaches can be applied for news articles having multiple locations specified in different context?

We study each of these research questions and present a workflow to capture locations that are facing algal bloom issue as reported by the news articles. We propose a probabilistic approach to classify the news articles to identify a specific category which confirms algal bloom occurrence at a specific location. We aim to extract this location information from the classified articles and generate a report of locations facing algal bloom for the CyanoTracker Web Application [7].

- For our study, we have preferred collecting news articles from Google News [8], an aggregator site which collects news articles from worldwide sources and presents them to the user based on his or her personalized interests. Our approach elaborates on the challenge of news articles extraction from the Google News [8] site. There are plenty of

technologies (like PHP, Python, JavaScript etc) to extract news articles. In our work we describe implementation of Google News Feeds API to extract news articles (in the form of feeds) related to the cyanobacteria bloom issue with the help of prominent keywords like "Blue Green Algae", "Cyanobacteria", "Red Tide", "Microcystin" and "CyanoHAB".

- We further describe the methodology we utilized to classify these extracted feeds to focus on the articles which confirm about algal bloom issue at a specified location. For this challenge, we utilize the unsupervised machine learning technique named Latent Dirichlet Allocation (LDA) [9]. LDA is a generative-probabilistic model for collection of discrete data such as text corpora which generates topic probabilities for each document in given text corpora. We present the results obtained after application of LDA using t-SNE 3D Plot [10], a dimensionality reduction technique devised to visualize high-dimensional data set into a space of two or three dimensions.

- We further elaborate on the implementation of Named Entity Recognition (NER) [11] framework to tag and recognize location keywords from each article classified into relevant category. NER is a popular natural language processing phenomena used to identify entities like Person, Location and Organization from given text. These tagged entities can be further used for various data analysis purposes or even as features for machine learning algorithms. In our work, we utilize these locations to track information about algae issue with the help of NER framework named Accurate Online Disambiguation of Named Entities in Text and Tables (AIDA) [12]. AIDA is a framework and online tool for entity detection and disambiguation which maps mentions of ambiguous names to canonical entities registered in the YAGO2 [13] knowledge base.

- We further elaborate on the multiple location keywords issue. For some news articles, the NER framework tags more than one location keyword per article thus creating the ambiguity as to which location is addressed primarily for algae bloom in given

article. We hence conclude this study, by illustrating on the multiple location keywords ambiguity challenge and then, formulate this challenge by using relative parameters like keyword frequency and news article headline to estimate the primary location addressed in a given article.

With our work, we hence suggest an effective approach to extract locations that are affected by algal bloom reported in news articles.

1.2 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, we describe the background and literature review of this project by discussing more about cyanobacteria and other web applications that are developed for cyanobacteria tracking. In Chapter 3, we discuss the motivation and overview by discussing the actual problem statement. In Chapters 4 and 5, we discuss related work and details of the approach that we have utilized for this work. Chapter 6 describes the results and validations. We conclude the thesis with Chapter 7 along with future work.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

In this section, we discuss the details of Cyanobacteria and its harmful effects. We also describe other web applications that are currently in use for algal bloom tracking.

2.1 Cyanobacteria

The harmful cyanobacteria are referred to by various names but more prominently such as "$BlueGreenAlgae$", "$Cyanobacteria$", "$CyanoHAB$", "$RedTide$" and "$Microcystin$". Hence we extensively use them in our data collection step to find articles that are addressing the cyanobacteria issue.

Cyanobacterial Harmful Algae Blooms (CyanoHABs) are harmful organisms growing on water bodies. These bacteria are capable of photosynthesis and can thrive in wide range of temperatures. They degrade the quality of water [14] . These harmful organisms grow on the surface of water bodies and are a major source of health hazards. They also degrade the quality of environment and cause hindrance on various recreational activities and destroy aquatic lives by killing fishes and also affecting humans and animals who are the consumers of this contaminated water [14]. They produce harmful toxins when found on the water bodies. These cyanotoxins, which are now part of the water bodies are hazardous to all living organisms that consume it directly or indirectly. Thus, the entire ecosystem gets affected by these cyanotoxins, putting large numbers of lives in danger [14]. CyanoHABs not only have an impact on the social and environmental life but also have significant economic impacts worldwide because of fish deaths and agricultural loss.

2.2 Traditional Approaches for Monitoring Cyanobacteria

The current cyanobacterial monitoring techniques include following approaches.

1. In-situ Monitoring
2. Satellite Sensing
3. Social Sensing

In-Situ methods involve intrusive approach where the field scientists, typically microbiologist collect water sample and analyze the sample under microscope or ELISA (enzyme-linked immunosorbent assay) kit [15]. This technique needs domain expertise or trained personnel who can follow the process carefully and thus the method gives high accuracy to the degree that cyanobacteria species type and the cell count can be identified because of the microscopic spatial resolution. However, it is time consuming, laborious, expensive considering the fact that visiting many lakes in a region will add up the cost and needs specialized domain training. At the same time it is not scalable to wider and remote geographic locations.

Satellite sensing exploits the optical nature of CyanoHAB captured through spectral sensors. Human eye can perceive information in the three visible bands (blue, green and red) of the electromagnetic (EM) wavelength. Unlike human eye, multi-spectral sensors are capable of diving the EM wavelength into multiple narrow bands and capture intensity of light reflected from each of these sub bands. Thus the plot of remote sensing reflectance vs wavelength (spectral profile) gives cyanoHAB a unique spectral profile shape that is easily distinguishable from other materials spectral profile. However, the multi-spectral sensors have coarse spatial resolution thus significantly compromising the accuracy. Also lakes that are smaller in size cannot be monitored with these sensors due to coarse spatial resolution. As opposed to multi-spectral sensor, hyper-spectral sensors captures information in contiguous wavelength. Although these hyper-spectral sensors are not able to identify the cyanoHAB

species, previous research has shown that the surface level cell concentration level using hyper-spectral sensors are comparable to in-situ analysis, thus they have high accuracy. However, their accuracy suffers when cyanobacteria is suspended few inches below the water surface.

With social media platforms like Facebook, Twitter and Instagram broadcasting information by allowing anyone to share information about oneself and their surroundings, large volumes of data has been made available. Exploitation of this information to analyze aspects of physical world is termed as Social Sensing. Many research challenges are focused on extraction, understanding and characterization of this data to generate valuable conclusion. The platforms are advantageous since they provide large volumes of data with geographically scattered users covering wide range of topics. On the contrary, validation and cleansing of this data are some of the challenges faced in this context.

2.3 CyanoTracker Web Application

CyanoTracker application [7] is an initiative by a group of researchers from Department of Computer Science and Department of Geography at The University of Georgia, for addressing the cyanobacterial bloom in the Georgia inland waters. CyanoTracker is unique in its data collection approach. It aims at integrating crowd sensed data (which is collected by the web application) with its remote sensors and social media platforms. The crowd sensed data is submitted by the users visiting a location and observing algae formations, in the form of images via the web or mobile application. The users can also submit their observations using projects email address or other social media platforms like Facebook, Twitter and Instagram.

For remote sensing, the project has sensors deployed at four lakes in Georgia (two in Lake Oconee and one each at Lake Oglethorpe and Lake Chapman - Sandy Creek Lake). It monitors these sensors closely to capture images of the bloom and integrate further with the satellite data. The application further encourages crowd to submit their observation about the blue green algae through images and descriptions with the help of the web or mobile

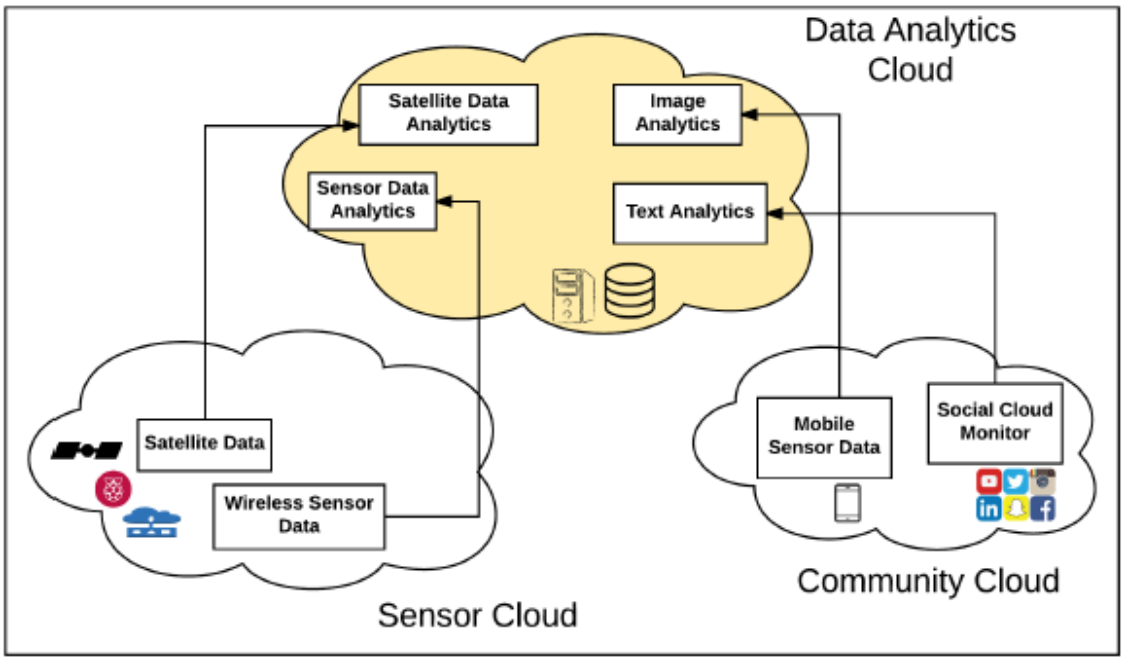


Figure. 2.1: CyanoTracker Application Architecture (Image Source - Dr. Vinay Kumar Boddula PhD Dissertation Thesis, Cyber-Social-Physical Approaches for Effective Detection of Cyanobacterial Blooms)

application. A large contribution to the applications content is through social media data. Twitter data analysis has been an ongoing research for this project. To improve the bloom reporting capture, the application recently began its data analysis for news articles through this research work.

2.4 Cyanobacteria Tracker

The Lake Champlain Committee (LCC) [16] in collaboration with Vermont Department of Health [17] have trained volunteers for capturing information on algal bloom. These volunteers periodically monitor the four lakes namely Lake Champlain, Lake Carmi, Lake Iroquois and Lake Memphremagog. The data collected from these volunteers is aggregated and uploaded in the web application named Cyanobacteria Tracker [18] and made available for the users.

2.5 HABSOS - Harmful Algal Bloom Observing System

Florida Fish and Wildlife Conservation Commission (FWC) [19] in collaboration with National Oceanic and Atmospheric Administration (NOAA) [20] has developed an application named Harmful Algal Bloom Observing System (HABSOS) [21]. The application primarily focuses on Gulf of Mexico for HAB information. It provides advanced information retrieval based on multiple parameters like species, date range and predefined conditions. The predefined conditions cover possibilities of water body closures, fish kills or detection by remote sensing. The results are displayed on a map view.

The FWC scientists combine field sampling with tools maintained by state and federal partners to track red tides and their effects. HABSOS application utilizes data provided by the FWC. The FWC works in close collaboration with University of South Florida, whose tools show size and location of blooms and forecast bloom movement.

2.6 CyAN - Cyanobacteria Assessment Network

This is a multi-agency project developed in collaboration by National Aeronautics and Space Administration (NASA) [22], National Oceanic and Atmospheric Administration (NOAA) [20] , U.S. Geological Survey (USGS) [23] , and Environmental Protection Agency (EPA) [24]. An early warning system that utilizes historical and current satellite data to detect algal blooms in freshwater within United States. It aims to refine their bloom detection algorithms for satellite platforms and identify landscape linkage causes of chlorophyll a and cyanobacteria blooms in freshwater systems [25].

2.7 BloomWatch

A Quality Assurance Program Plan (QAPP) by U.S Environmental Protection Agency for the Cyanobacteria Monitoring Collaborative Program [26] has led to the development of three applications namely BloomWatch, CyanoScope and CyanoMonitoring.

BloomWatch [26] is a mobile application which utilizes crowd-sourcing to find and report potential cyanobacterial blooms. It is a simple application available on both Android and IOS. With initial sign up, the application allows the user to get familiar with the appearance of algal bloom. The application allows to click a picture and records the user's submission to the project. It further allows for the user to provide any description. After the user completes his submission, the application sends this information to relevant state agencies.

2.8 CyanoScope

CyanoScope [26] is an application developed under the QAPP uses modern technologies and a social approach to learn more about cyanobacteria and their distribution. It uses social approach to identify the cyanobacteria present on the surface waters. As a future work, it specifies to utilize the submitted reports to map the spatial distribution and seasonal occurrence of cyanobacteria.

The application functionality starts with users collecting cyanobacteria with an instrument called tow. The user should prepare microscopic slides and identify cyanobacteria found in the sample. The user is then required to take pictures (microscopic view) of the sample found and submit it to iNatutalist.org [27], an online community resource that provides a platform for nature lovers to connect and record their observations.

2.9 CyanoMonitoring

CyanoMonitoring [26] is an application monitoring cyanobacteria in surface waters to determine the environmental factors that cause blooms. This is a long-term approach to understand how cyanobacteria respond to environmental conditions. The users must collect the water samples and analyze the samples for chlorophyll content using the equipment dedicated for this purpose. The website mentions about submission of their analysis as a future scope.

CHAPTER 3

MOTIVATION AND OVERVIEW

In this section we discuss limitations of the applications discussed in the Background and Literature Review chapter, the motivation of our research work and the research question that we are addressing.

Based on the applications discussed in the Background section, we analyze overall functionality characteristics. We observe field sampling by volunteers is a common data collection approach. Applications like CyAN (discussed in section 2.5) implement different technique like remote satellite sensing for data collection which utilizes current and historical satellite data. Other application, HABSOS (discussed in section 2.4) utilizes crowd-sourcing technique for data collection by allowing users to click a good quality image from the application and submit. For certain applications like CyanoScope (discussed in section 2.7), data sampling is done by crowd-sourcing but with an additional step of using dedicated instruments called tow for collecting samples. Post data collection, the applications either rely on image analysis, social media forums or relevant state agencies to confirm from the pictures submitted by the users or samples collected if cyanobacteria is present.

With all these applications we see the implementation of numerous techniques like crowd-sourcing, remote sensing and manual sampling for data collection. We also see social media approach to confirm cyanobacteria. But we observe that news articles have not been explored by any of the prior applications for algal bloom detection domain. This was our primary motivation to utilize this data source for gathering affected locations. With applications implementing manual field sampling, we understand the number of locations monitored

by these applications are limited. With news articles, we observed tracking locations from worldwide sources will be enabled. Our approach will hence gather affected locations from across the globe and will not be limited to any geographical boundaries, specialized instruments for sample collection or develop microscopic view pictures but instead rely on efficient machine learning techniques for text mining and analysis.

For our work, we specifically analyze news articles collected from Google News [8] with the help of its API called Google News Feeds to track locations affected by algal bloom. Google News is an aggregator site and hence a huge repository of news from worldwide sources. Numerous articles related to algal bloom at a specified location, fish kills, beaches closing, activity shut down on a lake are reported. By targeting specific articles which confirms bloom at a given location, we develop a workflow to extract location information from such articles. With our work we address following research questions:

- How can news articles be utilized to track algal bloom affected locations?
- How to classify news articles that report algal bloom at a location?
- How to extract the location information from the classified news articles?
- What location modelling approaches can be applied for news article containing multiple location keywords?

CHAPTER 4

RELATED WORK

In this section, we discuss in detail about the machine learning techniques and the named entity recognition frameworks that we utilized in our workflow.

4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation [9] is a generative probabilistic model for collections of discrete data such as text corpora. It is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document w in a corpus D:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n :
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

For simplified understanding, following figure depicts graphical representation of LDA.

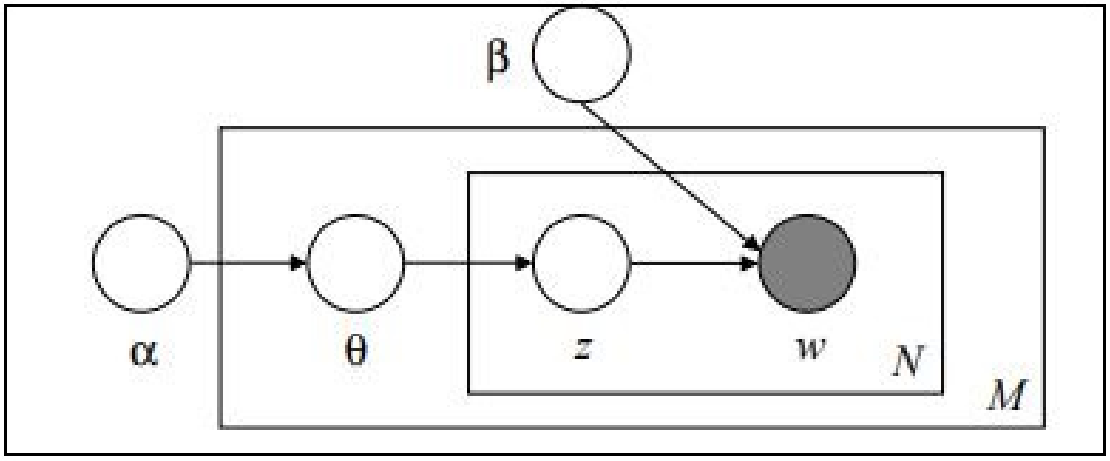


Figure. 4.1: Graphical Representation of LDA [9]

As the figure 4.1 makes it clear, there are three levels to the LDA representation. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document. For our implementation, we utilize sklearn's package named "Decomposition" which provides in built library LatentDirichletAllocation [28].

4.2 Named Entity Recognition

Named entity recognition (NER) [11] aims at identifying entities of interest in the text, such as person, location, organization and temporal expression. Identified entities can be used further in various downstream applications and information extraction systems. They can also be used as features for machine learning systems for other natural language processing tasks. Early NER systems relied on human defined rules. These rules based systems are time consuming and difficult to transfer to new types of entities. With the advent of semantic web knowledge base, the rules based systems for NER have been developed further. In our study, a framework which leverages this edge technology of semantic web to improve recognition.

For our initial analysis, we considered the Stanford NER [29]. It is a java based implementation of entity recognition developed by The Stanford Natural Language Processing Group. It provides classification for English language particularly for three class based classifier (Person, Organization and Location). It provides a maximum of seven classes tagging and hence the tagging accuracy for this framework was comparatively low compared to AIDA. For the news articles which specified names of certain lakes, the software could not tag accurately.

4.3 AIDA

AIDA is the acronym for Accurate Online Disambiguation of Named Entities in Text and Tables [12]. It is a framework and online tool for entity detection and disambiguation. Given a natural language text or table, it maps the occurrence of entities (namely People, Organization, Location) registered in the YAGO2 [13] knowledge base. YAGO is semantic knowledge base developed from Wikipedia [30] , WordNet [31] [32] [33] and GeoNames [34] . It has more than 10 million entities (like persons, locations, organizations) and contains more than 120 million facts about these entities. The accuracy of YAGO has been evaluated manually, providing an confirmed accuracy of 95%. The tagging ability of AIDA has been experimentally tested by annotation of 1400 newswire articles with entities mentioned in each article.

One of the AIDA's key contribution is creation of large, manually labelled ground truth data set to evaluate existing methods alongside theirs, comprising 1,393 documents and nearly 35,000 mentions. The annotations are publicly available, and many follow-up works have relied on this data. Along with accuracy and related information link provision, for a geographical entity AIDA has improvised recognition ability because its knowledge base tracks locations into large number of categories. It has 19 Geo-Entity categories defined which not only provisions it to easily recognize bigger entities like a country, state or national park but also smaller entities like garden or reservoir. Some examples of these categories are country, city, dam, desert, island, lake, river, temple and many more. The figure 4.2 describes

the types for all the entities that YAGO includes. AIDA service can be utilized either by installing the entity knowledge base on our server or online web service. For our study, we utilized AIDA through its web API.

person				artifact	event	organization	yagoGeoEntit
academician	businessman	pianist	senator	album	attack	agency	city
actor	businessperson	poet	singer	algorithm	battle	company	country
admiral	chemist	politician	soccer_player	artwork	catastrophe	enterprise	dam
adviser	coach	politician	socialist	book	conference	institute	desert
advocate	comedian	president	social_scientist	bridge	crime	league	garden
ambassador	composer	priest	songwriter	computer_game	conflict	party	hill
anthropologist	cricketer	princess	sovereign	device	election	team	island
architect	criminal	prince	spiritual_leader	memorial	festival	university	lake
aristocrat	cyclist	prisoner	swimmer	merchandise	invasion		monastery
artist	dancer	producer	teacher	motor_vehicle	revolution		mosque
astronaut	diplomat	professor	tennis_player	movie	riot		mountain
atheist	director	psychologist	theologian	musical	war		oasis
athlete	disk_jockey	reformer	trainer	musical_instrument			oilfield
aviator	doctor	ruler	vegetarian	painting			park
ballplayer	drummer	runner	violinist	publication			river
banker	duke	saint	wrestler	ship			temple
basketball_player	economist	scholar	writer	system			tower
biologist	editor	scientist					town
boxer	physicist	sculptor					village

Figure. 4.2: AIDA Framework Entity Types [12]

4.4 Naive Bayes

One approach to text classification is to assign to a given document d the class $c^* = argmax_c P(c|d)$. The Naive Bayes (NB) classifier is derived by first observing that by Bayes rule,

$$P(c|d) = P(c)P(d|c)/P(d),$$

where P(d) plays no role in selecting c^*. To estimate the term $P(d|c)$, Naive Bayes decomposes it by assuming the fi's are conditionally independent given d's class:

$$P_{NB}(c|d) := P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})/P(d)$$

The training method consists of relative-frequency estimation of $P(c)$ and $P(f_i|c)$, using add-one smoothing. Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well [35]. On the other hand, more sophisticated algorithms might (and often do) yield better results [36]. In our study, we implement multinomial naive bayesian model using sklearn library [28].

4.5 Support Vector Machine

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes [37]. They are large-margin, rather than probabilistic, classifiers, in contrast to Naive Bayes and MaxEnt. In the two-category case, the basic idea behind the training procedure is to find a hyperplane, represented by vector $\vec{w}$, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem; letting $c_j \in (1, -1)$ (corresponding to positive and negative) be the correct class of document d_j , the solution can be written as

$$\vec{w} := \sum_j \alpha_j c_j \vec{d_j}, \alpha_j \geq 0,$$

where the $\alpha_j's$ are obtained by solving a dual optimization problem. Those $\vec{d_j}$ such that α_j is greater than zero are called support vectors, since they are the only document vectors contributing to . Classification of test instances consists simply of determining which side of $\vec{w}'s$ hyperplane they fall on [36]. We implement support vector model using scikit-sklearn library [28].

4.6 Word Embeddings

Mikolov et al. introduced word2vec technique [38] that can obtain word vectors by training text corpus. The idea of word2vec (word embeddings) originated from the concept of distributed representation of words [39]. Word embeddings proved to be effective representations in the tasks of sentiment analysis [40] and text classification [41] [42]. We utilize the word2vec embedding by integrating it with both Multinomial Naive Bayesian and Support Vector Machine model.

4.7 Precision and Recall

Precision and Recall are the parameters used to measure accuracy of a classifier. It is a successful measure especially when the classes are imbalanced. Precision is the measure of

the result accuracy and Recall is the measure of how many truly relevant results are returned. High value of Precision implies low false positives and high value of Recall implies low false negatives. Hence high precision and recall values imply that the classifier is returning accurate results and majority of them are positive [28].

Precision (P) is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p).

$$P = T_p/(T_p + F_p)$$

Recall (R) is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n).

$$R = T_p/(T_p + F_n)$$

These quantities are also related to the (F_1) score, which is defined as the harmonic mean of precision and recall.

$$F_1 = 2 \times ((P \times R)/(P + R))$$

We use the above mentioned parameters to measure the performance of multinomial naive bayesian and support vector machine classifiers.

4.8 Google News

Google News is a news site that aggregates headlines from worldwide news sources. It is an up-to-date news source repository which aggregates all kinds of news, from all over the world, categorizes them in a defined hierarchy and presents to the users based on its optimized results presentation algorithms. It not just relies on Page Rank for display of news but additionally taps on its own unique ranking signals like user clicks, authority of a publication in given topic, freshness, geography and many more. It allows a user to customize the information he wants to see based on his topics of interest, trending news, location, publishers

etc. This customization feature simplified our data collection process to be restricted to the topic of algal bloom and cyanobacteria based on required keywords.

4.9 Google News Feeds

Google Alerts is notification service provided by Google. A service that sends user an email when it finds new results such as web pages, newspaper articles, blogs etc. that matches users search terms. Another improved way for delivering regularly changing web content was introduced by Google News soon after its establishment. It is called RSS Feeds - Rich Site Summary. Google News Feed is an API introduced by Google News for delivering its content in RSS format. The API enables a user to integrate RSS feeds with his website. Google Feed API allows Public Atom, RSS or Media RSS download using JavaScript. Python provides a simpler way to download results of Google News with universal feed parser library. Universal Feed Parser is a python module for downloading and parsing syndicated feeds. RSS 0.90, Netscape RSS 0.91, User-land RSS 0.91, RSS 0.92, RSS 0.93, RSS 0.94, RSS 1.0, RSS 2.0, Atom 0.3, Atom 0.1 and CDF Feeds are classified as syndicated feeds.

4.10 Why Google News?

As mentioned earlier, Google News provides coverage from worldwide sources. From prominent news sources like CNN, Fox News etc to local regional news agencies, all are covered by Google News. Hence the news coverage is widespread. It is possible to track news that is trending in any given part of the world. This will help us reach more locations.

The news articles cover a wide range of topics related to algal bloom. Not just the issue of algal bloom of a location but also related incidents like beach activity closure, fish mortality rate increase in certain areas, dog death due to choking on algae bloom, residents living near the shores complain about foul smell in the area etc. These varied incidents coverage helped us learn more about the severity of algal bloom issue through news articles.

No article sources are hidden. Google News provides all information about the original news source website, date-time when the article was published and author of the article. Hence verification of a source is possible if we are uncertain which makes these sources more reliable. Also the news agencies collaborate with state officials or authorities to confirm the to-be published content. Hence apart from source, the reliability for the article content is also more.

The article content is written by dedicated journalists who are meant to cover and distribute news to public. Hence the language of the content is letter-perfect. Unlike Twitter data it is not colloquial and contains less grammatical errors.

CHAPTER 5

LOCATION EXTRACTION WORKFLOW AND TECHNIQUE

The following section discusses our approach towards the articles classification and location modelling challenge and a detailed architecture description for the approach implementation.

5.1 Design Workflow

Figure 5.1 describes the step by step approach we utilize for article classification and location modelling. Our approach begins with extraction of news articles based on the prominent keywords such as "Cyanobacteria", "Blue Green Algae", "CyanoHAB", "Red Tide" and "Microcystin". We integrate Google News Feeds for news articles extraction using python's universal feed parser library. A news feed contains three basic parameters namely title, URL and description. A title is the header of the feed which is primarily the headline of the article. URL is the source web address and description is a summary of the feed. We extract these three feed parameters based on the keywords and load into MySQL database.

Using the source URL parameter we perform extraction of article contents. These contents are stored in the MySQL database. Post content extraction, we preprocess the data prior to classification. As a part of preprocessing we remove stop words, special characters, extra blank spaces, perform stemming and lemmatization.

The preprocessed articles are now ready to be used for unsupervised learning text classification using Latent Dirichlet Allocation technique (LDA) [9]. LDA is a probabilistic graphical modelling algorithm used to discover topics that are possibly present in given text corpus. Using LDA we generate two topics or clusters (like Relevant and Irrelevant categories) and validate the top keywords of both the clusters. We plot LDA results for a

better visualization using t-SNE 3D graph plot [10]. We observed one of the extracted topic by LDA has topic keywords that are more prominent in the reporting type of articles. Articles confirming algal bloom, giving warning alerts for a location, specifying incidents like fish and dog death, beach activity closure are considered as reporting type of news articles. We had originally began with article classification problem with an supervised learning approach by building a voting system using two primary supervised learning classifiers - Multinomial Naive Bayesian and Support Vector Machine. We also included modified versions of these two classifiers by adding word embedding as input. For this experiment we had labelled articles data which we utilize here for validation. We hence validate all the articles in both the clusters with our labels which we had generated previously for our supervised learning approach. Along with labels we also compare results of LDA with predictions of each of these classifiers for thorough validation. Once we validate LDA topics with our predefined labels and supervised learning techniques, we store IDs of articles in relevant cluster into MySQL database.

Our next step involves tagging location keywords from each reporting article. We perform this tagging using a prominent Named Entity Recognition framework called as AIDA - Accurate Online Disambiguation of Named Entities in Text and Tables [12]. AIDA is a framework and online tool for entity detection and disambiguation. Given a natural language text, it accurately maps mention of ambiguous names onto canonical entities based on YAGO2 knowledge base[13] . A semantic knowledge base derived from Wikipedia [30], GeoNames [34] and WordNet [31] [32] [33]. Using AIDA, we tag location keywords from each article of relevant cluster and generate location coordinates using Geo-location API [43].

As mentioned earlier, some articles along with the current location also tend to mention additional locations in an article that have been reported previously. In this case, for a single article AIDA tags multiple location keywords. For this issue, we utilize the article headline which most of the times contain the name of the primary location which the article is addressing to. Using the location mentioned in the headline, we can hypothesize that the

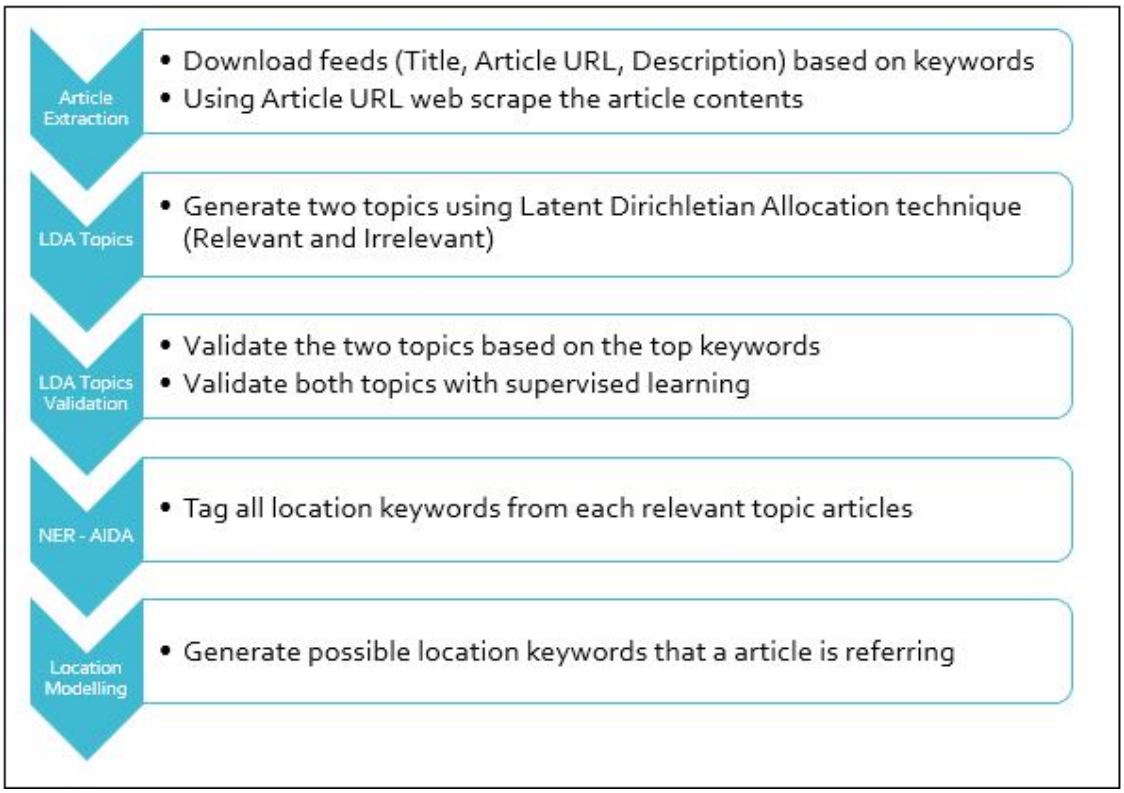


Figure. 5.1: Design Workflow for Article Classification and Location Modelling

article is primarily addressing to it than other tagged locations. Other intuition we consider is that an article will tend to repeat a location about which it is addressing currently more often than other locations. Since it is difficult to generate a definitive rule as to which locations can be disregarded, we generate a probabilistic output based on the location mentioned in headline and frequency of a location usage in the content.

5.2 Architecture

In this section, we elaborate each component of our design workflow. As seen in Figure 5.2, our architecture comprises of four main components namely Feeds Extract and Preprocess, Classification Model, Named Entity Recognition − AIDA and Geolocation coordinates. We further discuss the role of each component.

We begin with the search script within the feeds extract and preprocess component. It is a python script which is scheduled to run every hour and searches Google News site for availability of new articles based on our five prominent keywords. As and when the new articles arrive, the script checks if the article already exists in the database by computing Jaccard Similarity [44] between the description of each article in the database against the incoming article. If the similarity score is above the threshold then the new article is added to the database. Once all the new articles are added, the preprocessing script begins to clean the data by removing stop words, special characters, extra blank spaces and performs stemming and lemmatization. The processed content is stored in the MySQL database.

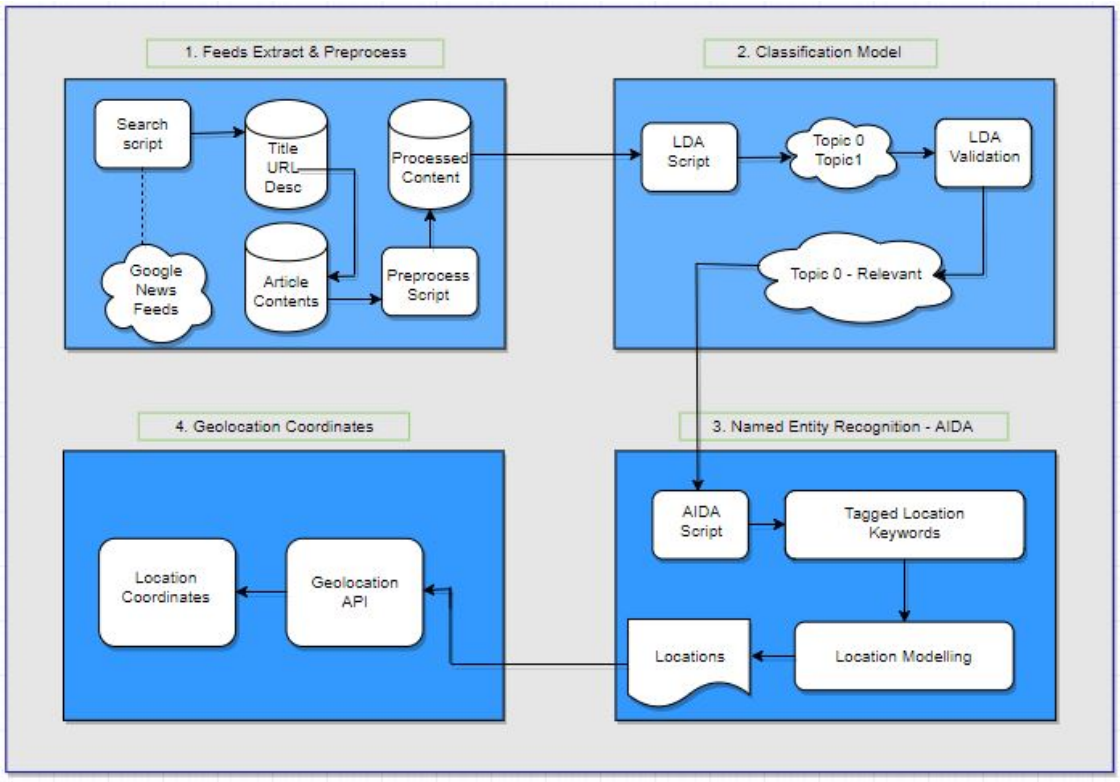


Figure. 5.2: Architecture Diagram

Here begins the role of component two i.e the classification model. We implement unsupervised learning technique named Latent Dirichlet Allocation (LDA) to generate two statistically separate topics, topic 0 (Relevant) and topic 1 (Irrelevant). On the initial observation over the topic keywords of both the clusters, we find top keywords of topic 0 to be

pointing towards the reporting articles. The second cluster that is the topic 1 has more random keywords in its top words list. We take a deeper look at these top words of both the generated topic in results section. To be thorough, we perform LDA validation initially by comparing the articles in both the topics with the prior labels which we had generated for supervised learning. We then further compare articles in both the topics with the predictions of supervised learning model. For supervised learning, we use Multinomial Nave Bayesian [45], Multinomial Nave Bayesian with Doc2Vec embedding, Support Vector Machine [46] and Support Vector Machines with Doc2Vec embedding. We compare predictions of each article from both the LDA clusters against all four supervised learning techniques.

After these validation, we ensure that topic 0 from LDA's cluster is most relevant. Next, we proceed to our third component which is Named Entity Recognition - AIDA. As mentioned previously AIDA is a framework and online tool which accurately disambiguates entities from given natural language text. We pass each relevant article through the AIDA framework and tag all the location keywords in each article. We then perform location modelling by comparing each tagged location with the location in the article headline and the frequency of that location keyword in the given article. Post location modelling we generate a probabilistic output for each article with its possible location keywords.

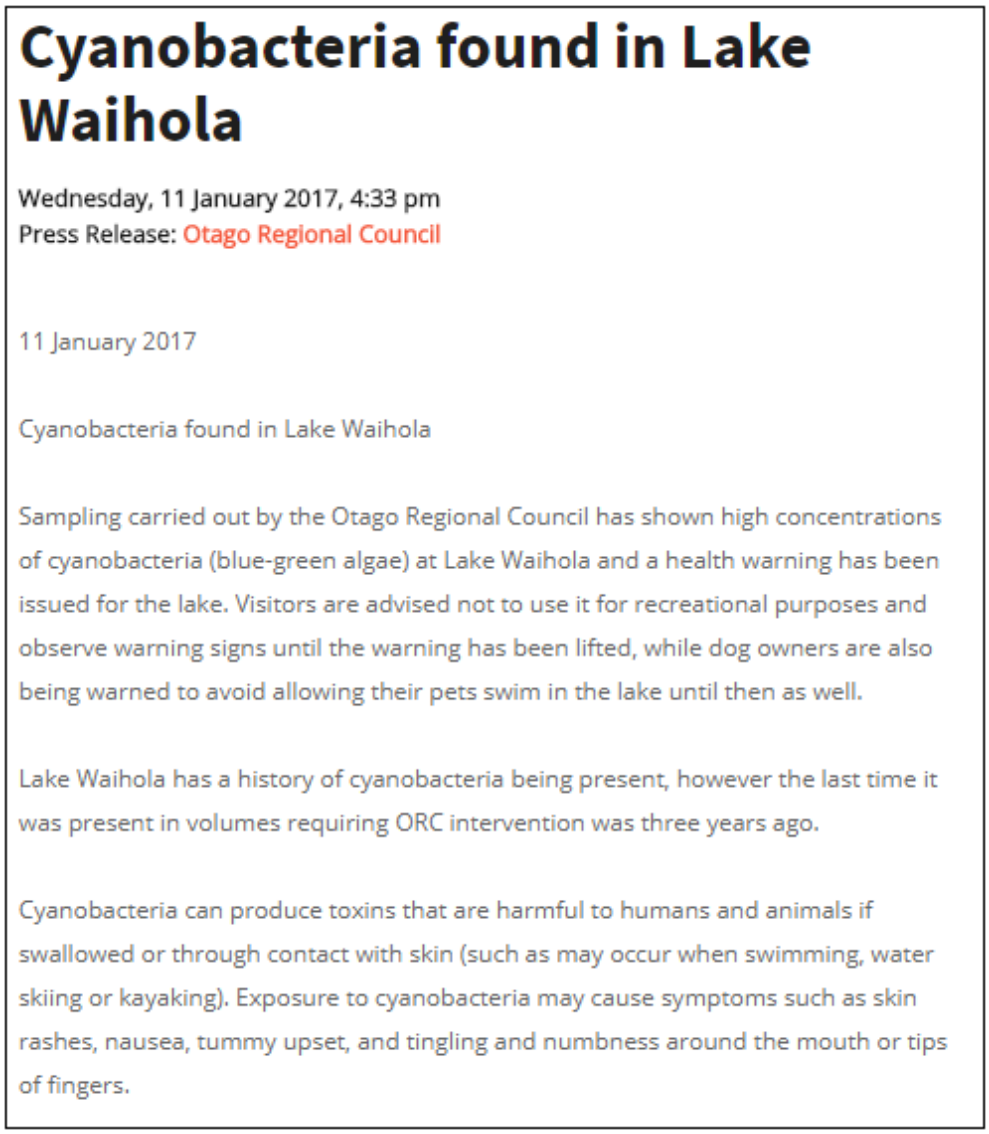
These location keywords are passed to the fourth component which is Geo-location Coordinates. It uses Google Maps Geo-location API [43] developed by Google to generate Geo-location coordinates based on the location name. We hence formulate the location keyword extraction from news articles with these components.

5.3 Location Modelling

A typical reporting article addressing a specific location begins with the algal bloom description at that location. It may additionally include name of an official regional entity who conducted required experiments to confirm the occurrence of bloom. The article may additionally mention other related information such as location is closed or swimming prohibited.

It may also include a brief overview of what cyanobacteria is, what precautionary measures should be taken if algae is seen and a contact to report about any other locations or incidents.

Figure 5.3 depicts a typical reporting article sample which addresses the bloom warning in Lake Waihola. The news is confirmed after sampling experiments conducted by the Otago Regional Council. It further mentions about a brief history of Lake Waihola in terms of algal bloom. The article ends on a note of general brief about cyanobacteria.



Cyanobacteria found in Lake Waihola

Wednesday, 11 January 2017, 4:33 pm
Press Release: Otago Regional Council

11 January 2017

Cyanobacteria found in Lake Waihola

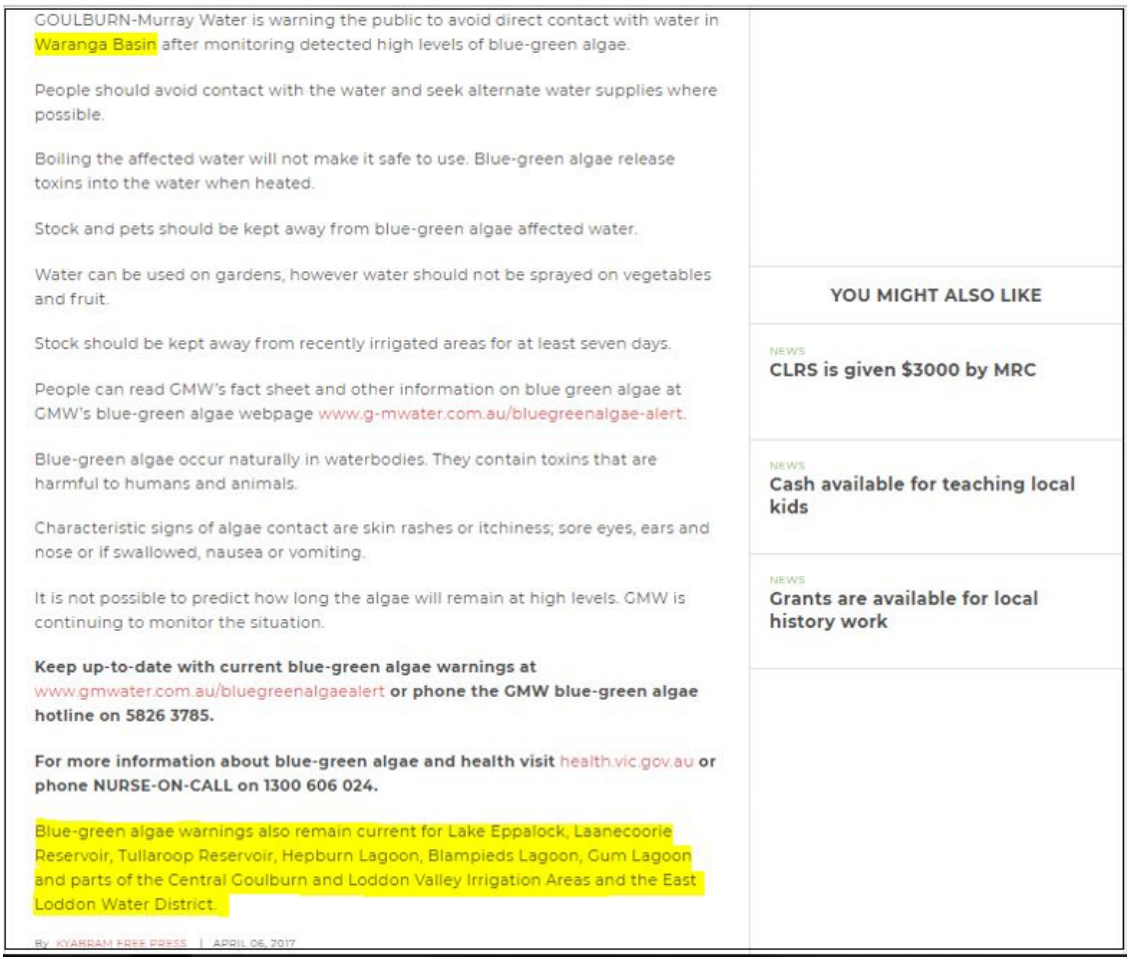
Sampling carried out by the Otago Regional Council has shown high concentrations of cyanobacteria (blue-green algae) at Lake Waihola and a health warning has been issued for the lake. Visitors are advised not to use it for recreational purposes and observe warning signs until the warning has been lifted, while dog owners are also being warned to avoid allowing their pets swim in the lake until then as well.

Lake Waihola has a history of cyanobacteria being present, however the last time it was present in volumes requiring ORC intervention was three years ago.

Cyanobacteria can produce toxins that are harmful to humans and animals if swallowed or through contact with skin (such as may occur when swimming, water skiing or kayaking). Exposure to cyanobacteria may cause symptoms such as skin rashes, nausea, tummy upset, and tingling and numbness around the mouth or tips of fingers.

Figure. 5.3: Sample Reporting Article with Single Location Keyword [1]

However certain articles tend to mention additional locations to address the severity of the issue in more depth, with the context of previously occurred reporting, possible bloom in future, currently warning issued. Figure 5.4 depicts an article which mentions multiple location keywords. The article is primarily addressing the algal bloom issue in Waranga Basin. However as seen in the figure 5.4, at the end of the article it lists other secondary locations that are also facing bloom issue including Lake Eppalock, Laanecoorie Reservoir, Tuliaroop Reservoir, Hepburn Lagoon, Blampeds Lagoon, Gum Lagoon, Central Goulburn, Loddon Valley Irrigation Areas and East Loddon Water District.



GOULBURN-Murray Water is warning the public to avoid direct contact with water in Waranga Basin after monitoring detected high levels of blue-green algae.

People should avoid contact with the water and seek alternate water supplies where possible.

Boiling the affected water will not make it safe to use. Blue-green algae release toxins into the water when heated.

Stock and pets should be kept away from blue-green algae affected water.

Water can be used on gardens, however water should not be sprayed on vegetables and fruit.

Stock should be kept away from recently irrigated areas for at least seven days.

People can read GMW's fact sheet and other information on blue green algae at GMW's blue-green algae webpage www.g-mwater.com.au/bluegreenalgae-alert.

Blue-green algae occur naturally in waterbodies. They contain toxins that are harmful to humans and animals.

Characteristic signs of algae contact are skin rashes or itchiness; sore eyes, ears and nose or if swallowed, nausea or vomiting.

It is not possible to predict how long the algae will remain at high levels. GMW is continuing to monitor the situation.

Keep up-to-date with current blue-green algae warnings at www.gmwater.com.au/bluegreenalgaealert or phone the GMW blue-green algae hotline on 5826 3785.

For more information about blue-green algae and health visit health.vic.gov.au or phone NURSE-ON-CALL on 1300 606 024.

Blue-green algae warnings also remain current for Lake Eppalock, Laanecoorie Reservoir, Tullaroop Reservoir, Hepburn Lagoon, Blampieds Lagoon, Gum Lagoon and parts of the Central Goulburn and Loddon Valley Irrigation Areas and the East Loddon Water District.

By KYABRAM FREE PRESS | APRIL 06, 2017

YOU MIGHT ALSO LIKE

NEWS
CLRS is given $3000 by MRC

NEWS
Cash available for teaching local kids

NEWS
Grants are available for local history work

Figure. 5.4: Sample Article with Multiple Location Keywords [2]

Hence when we screen this article from AIDA, it recognizes all the location keywords generating multiple locations for given article. Hence resolving between all the recognized

locations and detecting which keyword is the one which the article is addressing primarily is our second challenge.

We first approached this challenge as a classification problem. Our first instinctive experiment was application of LDA but with more number of output clusters to be generated. The idea behind LDA application was to verify if any of the generated clusters prominently included location keywords. By manual analysis of the 660 articles from relevant cluster generated by LDA in previous step, we discovered a total of 108 unique locations were expected. We hence generated 108 clusters from LDA and analyzed the top keywords for each of them. None of the clusters had prominent location keywords and we knew this challenge will require another approach.

To overcome the challenges of the naive approach we next considered supervised learning for a sentence level classification. We labelled each article having multiple location keywords with primary location as ground truth and provided these labelled articles to a supervised learning based classifier. Majority of the labels were unique and hence we anticipate the learning process will not be efficient. Besides with every new incoming article, human intervention would be required to provide appropriate label. We hence discarded this approach as well.

We now approached the challenge with very basic human intuition. An article might tend to repeat the usage of the location keyword which it is addressing primarily in the article than other locations which it mentions in different context. In the example article presented in figure 5.3, the article specifies Lake Waihola twice. We exploit this observation and record the occurrence frequency of every recognized location keyword for each article. For resolution amongst all the identified locations from a given article, we first check this frequency of occurrence for each keyword. If amongst all the keywords, any one particularly occurs more frequently than others then we estimate that location to be the primary addressed one.

But for certain articles, the occurrence frequency did not help in complete resolution. We came across articles with multiple keywords and each occurring just once. As seen in the article specified in Figure 5.4, there are multiple location keywords but all occur only once including the primary addressed location. In this situation, we will require new parameter to validate. With our articles observations, we realized that the article headline occasionally contains the primary addressed location keywords. This fact can be utilized in scenario of single occurrence of all keywords. Hence for articles having multiple location keywords with singular frequency, we resolve the ambiguity by checking if the headline contains one of the keyword.

Another approach we researched included preservation of context for each location keyword. For locations which were mentioned in the context of previously reported, we observed usage of specific keywords like 'earlier','before', 'last week' and similar more. If for each keyword we preserve a list of time line specific keywords and filter based on the one that have present tense context. But this seemed very definitive rule and targeting only a few of all the keywords. For every incoming new context we will have to define a new definite rule. Hence we disregarded this approach and for the simplicity of this task we proceeded only by combining our intuition of occurrence frequency with article headline and generate a probabilistic estimate than a definitive filtered output.

CHAPTER 6

RESULTS

The following section describes the results for each of the component discussed in the architecture section. For experiments and evaluation, we are using a total of 1374 articles for both LDA and supervised learning validation. Following figure 6.1 shows the data distribution.

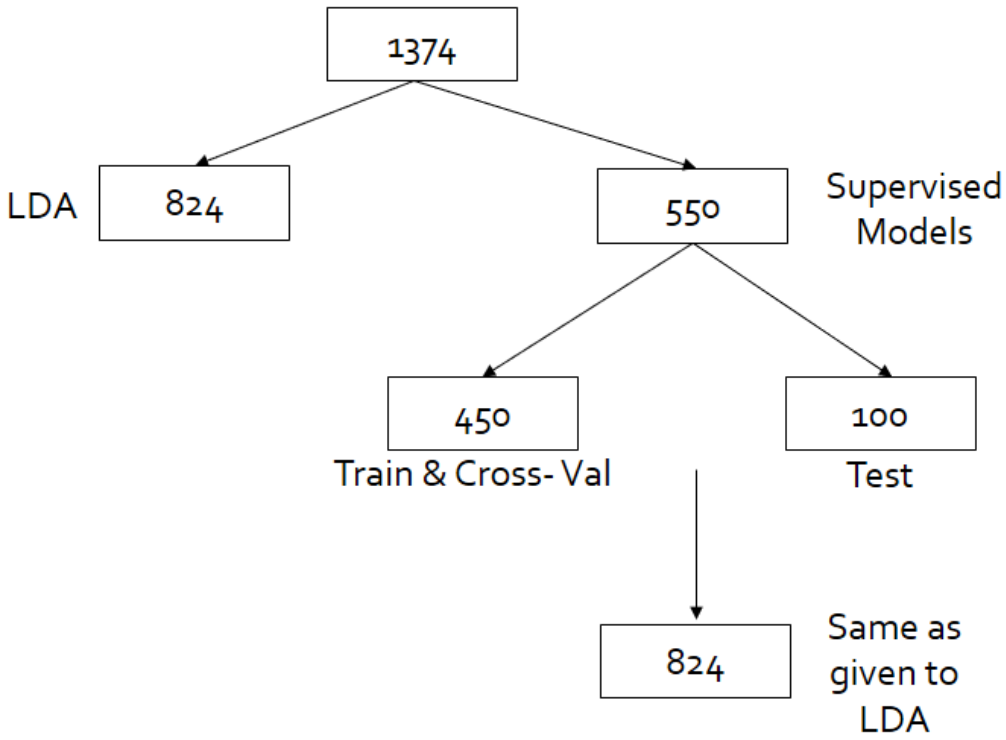


Figure. 6.1: Data Distribution

As seen in figure 6.1, we have a total of 1374 articles. Out of these 1374 articles, we utilize 824 for LDA classification. LDA being unsupervised learning technique, we would not have to divide these 824 into train or test set. The remaining 550 from 1374 articles

are utilized to train and prepare our supervised learning models. We hence divide the 550 articles as 450 training and cross-validation set and 100 as test set. Based on the accuracy for test set of 100 articles we decide if the model is ready and trained enough to now predict on unseen data set of same 824 articles which we utilized for LDA. We provide the same 824 articles to supervised models since we want to compare LDA predictions accuracy with supervised learning predication accuracy.

6.1 LDA Topics

The preprocessed 824 articles are the input for our LDA script. It generates two topics with the following top keywords.

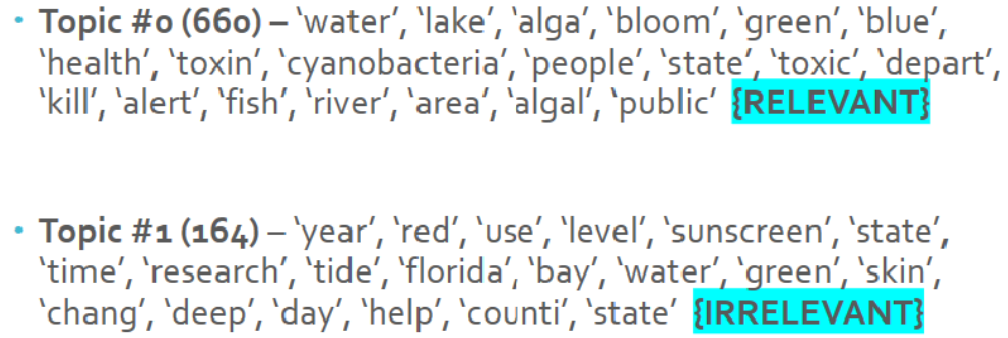


Figure. 6.2: Top Keywords from LDA Generated Topics

With analysis of the top words from both the topics, we understand topic 0 to be of more relevance compared to topic 1. Words like water, algae, bloom are weighed more in topic 0. The word 'depart' corresponds to department and in most of the reporting articles we had observed mention of an authority or health department of a location which speaks about algal bloom. There were couple of warning articles in our labelled list which specified about certain lakes are on alert and warning signs have been raised. We see a match for this scenario articles with the keyword 'alert' in topic 0. Some articles mentioned about many dead fishes floating on the shore due to the bloom. Topic 0 tends to cover these articles with its keyword 'kill'. Hence with top keywords analysis, we conclude topic 0 to be relevant. We

further investigate by comparing every article from both the topics with corresponding labels and supervised learning algorithms predictions.

- Relevant: 0.425* " water " + 0.399* " lake " + 0.347* " alga " + 0.301* " bloom " + 0.299* " green " + 0.2938* " blue " + 0.226* " health " + 0.224* " toxin " + 0.219* " cyanobacteria " + 0.216* " people " + 0.210* " state " +0.205* " toxic " + 0.198* " depart " + 0.183* " kill " + 0.167* " alert " + 0.161* " fish " + 0.153* " river " + 0.146* " area " + 0.132* " algal " + 0.123* " public "

- Irrelevant: 0.429* " year " + 0.276* " red " + 0.261* " use " + 0.259* " level " + 0.251* " sunscreen " + 0.244* " state " + 0.239* " time " + 0.233* " research " +0.222* " tide " +0.217* " florida " + 0.207 * " bay " + 0.201 * " water " + 0.198 * " green " + -0.188* " skin " + 0.163* " chang " + 0.158* " deep " + 0.153* " day " + 0.148 * " help " + 0.139* " counti " + 0.133* " state "

Figure. 6.3: Word Weight Distribution Per Topic

Figure 6.3 depicts the weights corresponding to each word in each topic. LDA not only generates topic distribution for each document passed through it but also word distribution for each topic. We generate two topics using LDA and we first analyze the word distribution for each topic. Hence we look at the top 20 words assigned to each topic in the order of their assigned weights by LDA. We observe some words like water and lake are assigned in both the topics but the corresponding weights for them are different and comparatively higher in case of Relevant topic. Negative weight terms are thematically similar to the topic, but not directly relevant. With this word weight distribution, we observe the topic labelled as Relevant to be of more relevance since with keywords like kill, alert and fish it focuses on articles that were reporting specific incidents. Topic labelled as Irrelevant in figure 6.3, has one keyword "sunscreen" which we came across in articles which addressed using cyanobacteria for creation of sunscreens for skin. In our labels these articles were Irrelevant. Hence with the word distribution analysis we hypothesize that topic 1 is pointing more towards

irrelevant articles as compared to topic 0 keywords. We further plot the topics generated by LDA using a high-dimensional data visualization tool.

6.2 LDA t-SNE 3D Plot

t-Distributed Stochastic Neighbor Embedding (t-SNE) [10] is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high dimensional data. We employ the t-SNE technique to plot our LDA topics results by providing the LDA generated probability output as an input to the t-SNE plot model.

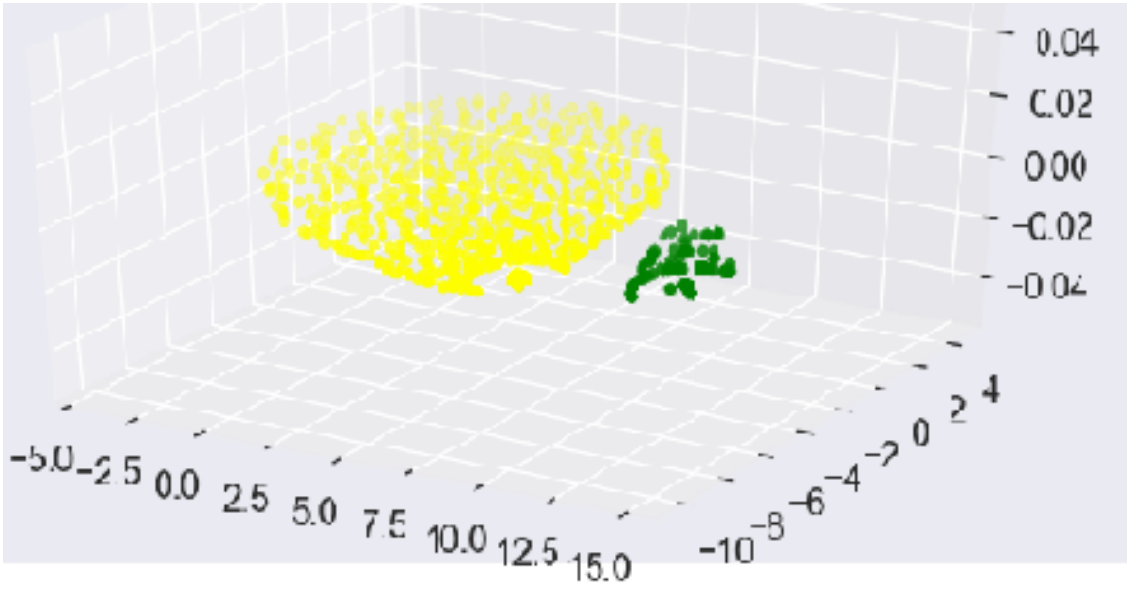


Figure. 6.4: LDA t-SNE 3D Plot for Topic Extraction from Articles

The yellow cluster is of topic 0 and the green cluster is of topic 1. Out of the total 824 articles, 660 were categorized into topic 0 and 164 in topic 1. Output of LDA is a list of length equal to the total number of articles (in this case 824) where each element corresponds to a sub-list containing probabilities each article being in topic 0 as well as topic 1. For example [[0.99185699 0.00814301], [0.10907101 0.89092899], [0.64845976 0.35154024]..., [0.86533694 0.13466306], [0.84888982 0.15111018], [0.63311712 0.36688288]]; where 0.99185699 is the

probability of article 1 being in topic 0 and 0.00814301 is the probability of article 1 being in topic 1. We provide this output of LDA in the similar format to t-SNE visualization where the X axis is the probability of article being in topic 0 and Y axis is the probability of article being in topic 1 which generates the cluster shown in figure 6.4.

6.3 LDA Validation

We validate each article from both the topics generated by LDA with the labels we had assigned as well as with supervised learning based classification. The table below discusses the number of labelled relevant and irrelevant for articles in both the LDA topics. In topic 0, LDA has generated total of 660 articles. We check the labels for all of these articles and find 648 articles labelled as relevant. 12 Irrelevant labelled articles were miss-categorized by LDA as Relevant. Similarly we check the labels for topic 1 articles and find out of 164 we had labelled 147 articles as Irrelevant. 17 Irrelevant articles were miss-categorized by LDA as Relevant. Hence based on the labels comparison, we can be assured that LDA articles are semantically apart enough to focus on articles reporting locations.

Table. 6.1: Validation with Labels

Classifier	LDA	Label Relevant	Label Irrelevant
topic 0	660	648	12
topic 1	164	17	147

For supervised learning, we model a Multinomial Naive Bayesian [45] and Support Vector Machine [46] classifier. We additionally include word embedding (labelled as D2V meaning document to vector) with both the classifiers and record the predictions. We now consider the new set comprising of 550 articles (338 − Relevant and 212 − Irrelevant) for supervised learning validation task. We utilize 450 articles for training, cross-validation and 100 articles for testing. Once the test set accuracy is optimal, we use these models to predict

on the 824 articles that we had earlier passed to LDA. The table below depicts the accuracy measures for cross-validation on the set of 450 articles.

Table. 6.2: 5-fold Cross-Validation Accuracy

Classifier	Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Average
SVM	0.943	0.961	0.963	0.947	0.948	0.95
SVM-D2V	0.849	0.866	0.843	0.844	0.865	0.85
MNB-D2V	0.820	0.846	0.829	0.824	0.851	0.83
MNB	0.752	0.759	0.781	0.786	0.784	0.77

As seen in Table 6.2, SVM has the optimal cross-validation accuracy. Multinomial Naive Bayesian classifier comparatively has better accuracy with word embeddings.We implement SVM with C as 1.0, gamma as 0.7 and kernel as rbf. We now note the precision and accuracy measures for the test set comprising of 100 articles.

Table. 6.3: Precision and Recall Measures for Test Set

Classifier-Measure	T_p	F_p	F_n	P	R	F_1
SVM	87	11	2	0.88	0.97	0.92
SVM-D2V	77	16	7	0.82	0.91	0.86
MNB-D2V	70	21	9	0.76	0.88	0.81
MNB	66	23	11	0.74	0.85	0.79

As seen in Table 6.3, the models provide us with considerable accuracy on the test set. SVM again gives us better accuracy as compared to naive bayes. We now record the performance for these models on the unseen 824 articles.

As we see in Table 6.4, SVM performs well with highest number of true positives. Not only high true positive but comparatively lower false positives and false negatives as well. CyanoTracker being an early warning system, lower false positives and false negatives are required. SVM with word embedding has good number of true positives. The number of negatives are comparatively high. We hence see articles that are in general discussing about cyanobacteria are classified as relevant. In comparison of MNB and MNB with word embedding, the word embedding version works comparatively well. But for both versions of naive bayes, both false positives and false negatives are extremely high. Hence for further processing we more rely on SVM.

Table. 6.4: Precision and Recall Measures for 824 articles

Classifier-Measure	T_p	F_p	F_n	P	R	F_1
SVM	806	13	5	0.98	0.99	0.98
SVM-D2V	758	49	17	0.93	0.97	0.94
MNB-D2V	645	60	119	0.91	0.84	0.91
MNB	618	66	140	0.90	0.81	0.85

6.4 AIDA Tagging

After LDA validation with supervised learning models, we see SVM gives most reliable accuracy. We hence see the number of articles from topic 0 (660 articles) classified by LDA, and predicted as relevant by SVM as well.

Following figure 6.5 discusses the number of articles we begin with and choose to move forward for AIDA recognition. As seen in the figure 6.5, we begin with 1374 articles, out of which 550 we preserve for building our supervised models. The remaining 824 articles are first passed to LDA and two topics (topic 0 − Relevant and topic 1 − Irrelevant) are generated. Before beginning the validation we prepare our supervised learning models with

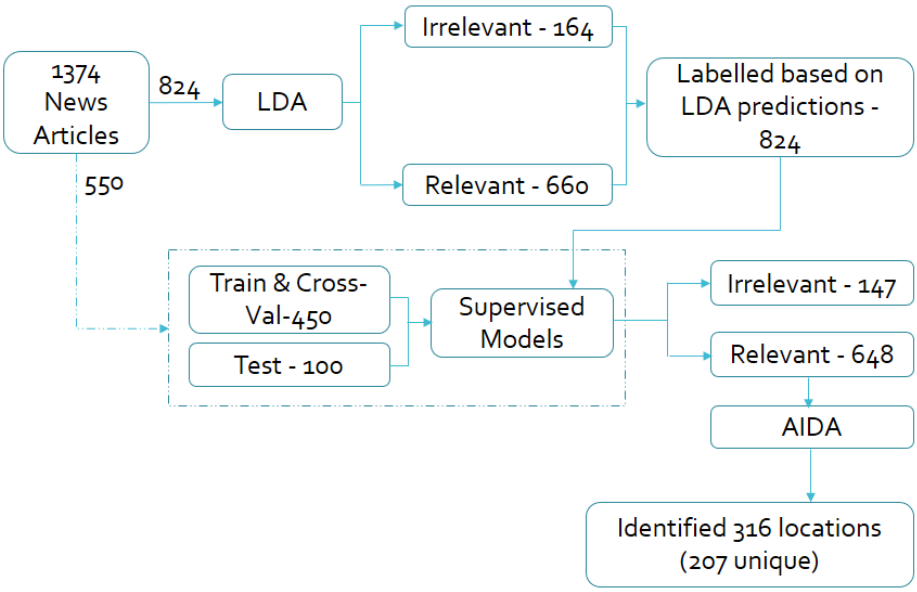


Figure. 6.5: Total Identified Locations from 824 articles

550 articles that were hand labelled. Out of the 550, we train and cross-validate on 450 articles and further test on 100. We record these performance to verify if the models are trained well enough to now predict on the unseen 824 articles.

We now test the supervised models on the 824 articles, same set that we initially passed to LDA for generation of two topics. Since SVM provides us the best performance, we focus on its predictions and understand that out of the 660 relevant predicted by LDA, 648 were predicted relevant by SVM as well. The remaining 12 from topic 0 were predicted Irrelevant. We hence consider only the 648 articles for further processing by AIDA. From these 648 articles, AIDA recognizes a total of 316 locations. Some of the locations had multiple articles addressing its issue. Hence a total of 207 unique locations were identified. From our work hence address the time frame and countries for these 207 locations.

Table. 6.5: Locations Tagged by AIDA for Articles from Relevant Topics Cluster

Article ID	AIDA Recognized Keywords
55	Utah, Utah Lake, Mississippi River, Lake Erie, Jordan River, Salt Lake City
152	Sacramento, San Luis Reservoir, Merced County
225	Upper Klamath, Klamath River, California, U.S, Lake Ewauna, J.C Boyle Reservoir, Copco Reservoir
368	Lake Needwood, Lake Frank
507	Mormon Reservoir, Camas County

Table 6.5 depicts some of the locations from the list of total 207 that were identified by AIDA within given articles. Some articles have multiple location keywords. Hence we need to model the addressed location for a given article with the help of modelling parameters - title and keyword frequency.

6.5 Location Modelling

As seen in the table 6.5, each article has multiple location keywords identified by AIDA. We hence apply our headline and word frequency based intuition techniques to predict which

location keyword is more relevant. For both these articles we check the headline and word frequency for location keyword.

Table. 6.6: Estimated Locations After Application of Location Modelling Intuitions

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
55	Toxic algae bloom closes Utah lake, sickens more than 100 people - Fox News	Utah - 2, Utah Lake - 4, Mississippi River - 1, Lake Erie - 1, Jordan River - 1, Salt Lake City - 1	Utah Lake
152	Blue-green algae bloom in San Luis Reservoir and O'Neill Forebay; caution urged in water contact - Lake County News	Sacramento - 1, San Luis Reservoir - 6, Merced County - 1	San Luis Reservoir
225	High Toxin Levels And Algal Blooms Found In Klamath Waterways - Tech Times	Upper Klamath - 2, Klamath River - 3, California - 2, U.S -1, Lake Ewauna - 1, J.C Boyle Reservoir - 2, Copco Reservoir - 1	Klamath River
		Continued on next page	

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
368	Montgomery Parks Reports Elevated Levels of Microcystin at Lake Needwood and Lake Frank - Montgomery Community Media	Lake Needwood - 1, Lake Frank - 1	Lake Needwood, Lake Frank
507	Health officials warn of blue-green algae at Mormon Reservoir - Twin Falls Times-News	Mormon Reservoir - 6, Camas County - 1	Mormon Reservoir

As seen in table 6.6, we thereby observe the headline and word frequency for each tagged location from all the articles. For article 55, 152 and 507, the keyword occurrence frequency itself fairly helps us make a clear decision. In all the three articles, a given keyword occurs very frequently as compared to other keywords. In article 55, Utah Lake occurs four times while other recognized keywords occur just once. Similarly, keyword San Luis Reservoir occurs six times as compared to single occurrence of other keywords. The keyword Mormon Reservoir also prominently stands out in article 507 by occurring six times as compared to single occurrence of other keywords. We further verify the estimated keywords (Utah Lake, San Luis Reservoir and Mormon Reservoir) for their occurrence in the respective article headline. Since the keywords are mentioned in their respective headline, we estimate these three would be the primary addressed keywords in the respective articles.

In articles 225 and 368, the word frequency does not help us much to estimate since the count is fairly same for many keywords. Article 225 has Klamath River occurring three times, but other keywords like Upper Klamath, California, J.C Boyle Reservoir occurring with similar frequency of two. Hence it is tricky to make an estimation here based on the word frequency. We hence consider the location name mentioned in the headline that is Klamath. Hence based on combined analysis, we estimate Klamath river.

Similarly article 368 has two keywords, Lake Needwood and Lake Frank occurring once. Based on the frequency it is difficult to estimate and hence we consider locations in article headline. The article headline contains both these locations and hence we estimate both Lake Needwood and Lake Frank to be the primary addressed locations.

6.6 Month Wise Reporting for Year 2017

For our research, we have captured data from August 2016 to December 2017. We present the monthly recordings for the year 2017 with following graph.

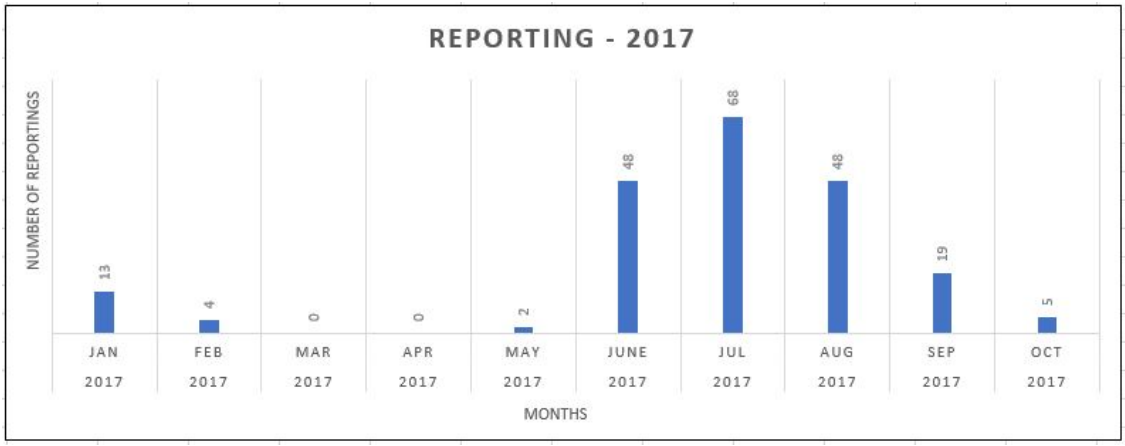


Figure. 6.6: Month Wise Reporting for the Year 2017

As seen in Figure 6.6, the year began with substantial number of reporting in January and further decreasing the number in February. Summer is the best period for the cyanobacteria to thrive and hence we see a spike in the number of reporting beginning from the month

of May. July of the year 2017 has seen the highest number of reporting from worldwide locations. With the decrease in temperature, growth of cyanobacteria is restricted and hence we see a drop in the number by the end of the year.

6.7 Country Wise Reporting for Year 2017

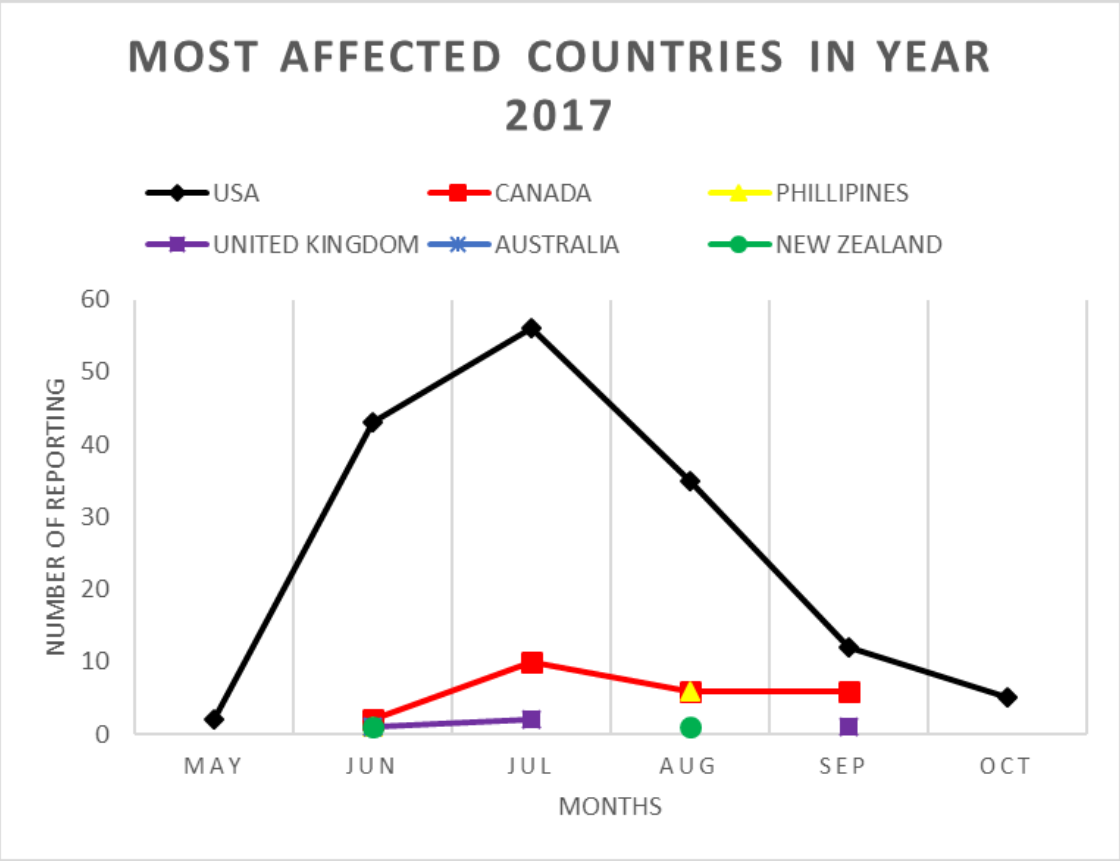


Figure. 6.7: Country Wise Reporting for the Year 2017

Figure 6.7 depicts the country wise reporting for the year 2017. Since May to October 2017 has seen a series of reporting, we capture the locations that were primarily reported within this period. From our observations, we see five prominent countries facing bloom from a range of high to low concentration are United States, Canada, Philippines, United Kingdom, Australia and New Zealand (in the order of severity). United States has seen the highest bloom amongst all affecting more than 150 water bodies. Canada has seen the second highest bloom affecting more than 25 water bodies. The issue is beginning from this year

for Philippines, United Kingdom, Australia and New Zealand with 7, 4 and 1 each reporting respectively. With future attention to the reported water bodies, these countries can avoid the bloom from growing further.

6.8 Google News Articles Reported Locations in 2017

A total of 108 locations were detected by our approach from the relevant topic cluster including a total of 660 reporting articles. For the sake of space, we present 22 locations detected in table 6.7.

Table. 6.7: Affected Locations from 2017 Google News Articles

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
55	Toxic algae bloom closes Utah lake, sickens more than 100 people - Fox News	Utah - 2, Utah Lake - 4, Mississippi River - 1, Lake Erie - 1, Jordan River - 1, Salt Lake City - 1	Utah Lake
152	Blue-green algae bloom in San Luis Reservoir and O'Neill Forebay; caution urged in water contact - Lake County News	Sacramento - 1, San Luis Reservoir - 6, Merced County - 1	San Luis Reservoir

Continued on next page

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
225	High Toxin Levels And Algal Blooms Found In Klamath Waterways - Tech Times	Upper Klamath - 2, Klamath River - 3, California - 2, U.S -1, Lake Ewauna - 1, J.C Boyle Reservoir - 2, Copco Reservoir - 1	Klamath River
368	Montgomery Parks Reports Elevated Levels of Microcystin at Lake Needwood and Lake Frank - Montgomery Community Media	Lake Needwood - 1, Lake Frank - 1	Lake Needwood, Lake Frank
507	Health officials warn of blue-green algae at Mormon Reservoir - Twin Falls Times-News	Mormon Reservoir - 6, Camas County - 1	Mormon Reservoir
510	Toxic blue-green algal bloom found at Magic Reservoir - KMVT	Idaho - 1, Magic Reservoir - 2, Blaine County - 1, Lava Point - 1	Magic Reservoir

Continued on next page

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
789	Toxic blue-green algae found at Summit Lake - The Daily World	Summit Lake - 1, Thurston County - 1	Summit Lake
825	Brookville Lake Under Blue-Green Algae Advisory - Eagle 99.3 FM WSCH	Brookville Lake - 1, Indiana - 1	Brookville Lake
833	Blue-Green Algae Detected in Wainscott Pond - East Hampton Star	Wainscott Pond - 3, Georgica Pond - 2, Wickapogue Pond - 2	Wainscott Pond, Georgica Pond, Wickapogue Pond
909	Feds, state close Marion Reservoir because of blue-green algae - Wichita Eagle	Marion Reservoir - 1, Wichita - 1, Marion County Lake - 1	Marion Reservoir
922	Advisory issued after toxic algae found at Lake Tapps - KOMO News	Lake Tapps - 1	Lake Tapps
995	Blue-green algae confirmed in Windermere - The Westmorland Gazette	Windermere - 1, Millerground - 1, Rayrigg Meadow - 1	Windermere

Continued on next page

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
1021	Algal blooms already causing beach closures in Dane County - WKOW 27: Madison, WI Breaking News, Weather and ... - WKOW	Dane County - 1, Lake Monona - 2, Wisconsin - 1	Lake Monona
1022	High levels of toxic blue-green algae close lake at Washington State ... - The Olympian	Anderson Lake State Park - 2, Washington State Park - 2, Summit Lake - 1, Thurston County - 1	Anderson Lake State Park, Washington State Park
1026	Bacteria Warning For Silver Lake In Hollis - Patch.com	Hollis - 2, New Hampshire - 1, Silver Lake - 2, Griffin Beach - 2, Webster Lake - 1	Silver Lake
1047	Toxic algal bloom in Vasse Estuary - Busselton Dunsborough Mail	Vasse River - 2, Busselton - 1	Vasse River

Continued on next page

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
1064	Toxic blue-green algae bloom stretches across Lake Mendota & Yah - WKOW 27: Madison, WI Breaking News ... - WKOW	Lake Mendota - 1, Yahara River - 2, Tenny Park - 2	Lake Mendota, Yahara River
1084	Buckeye Lake under algal bloom advisory - Fox 28	Buckeye Lake - 1, Columbus - 1, Ohio - 1	Buckeye Lake
1088	Cyanobacteria Advisory Issued For Elm Brook Park - Patch.com	Hopkinton Everett Lake -1, Elm Brook Park Beach - 2	Elm Brook Park Beach
1099	Blue-green algae found in northern part of Grand Lake - Four States Homepage	Oklahoma - 1, Grand Lake - 3, Fly Creek - 2	Grand Lake
1104	Blue-green algae warning issued for parts of Lake Isle - CBC.ca	Lake Isle - 2, Edmonton - 1	Lake Isle

Continued on next page

Article ID	Headline	AIDA Recognized With Frequency	Estimated Locations
1111	Do not use the water: blue-green algae found in Whalley Lake ... - NorthBayNipissing.com	Whalley Lake - 1	Whalley Lake

6.9 Satellite Imagery Integration

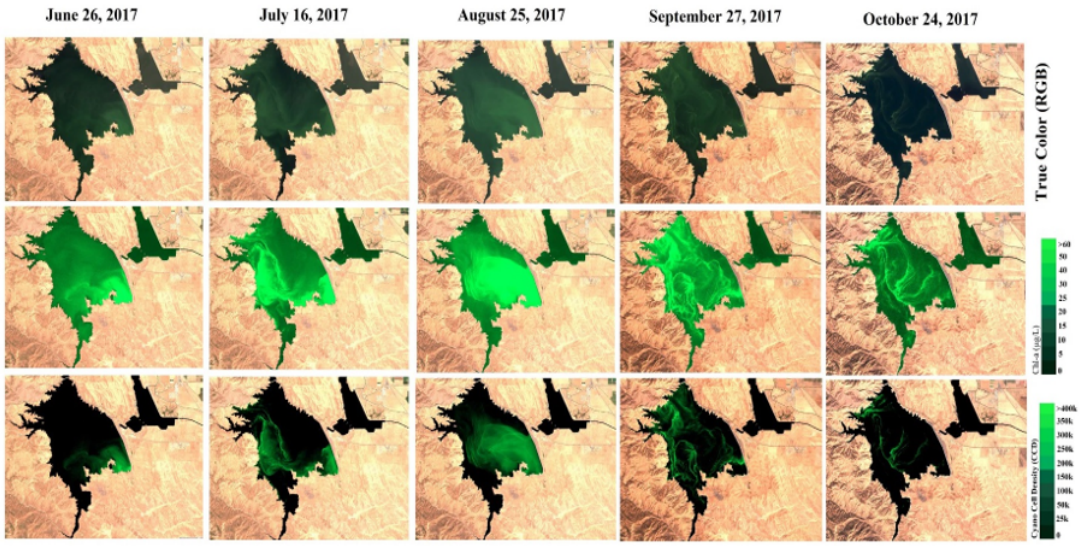


Figure. 6.8: Satellite Capture of Bloom Movement for San Luis Reservoir, California Reported in News Articles on June 21, 2017 [7]

Figure 6.8 depicts location San Luis Reservoir in California that was identified for cyanobacterial bloom on 21st June, 2017. We collaborate with Geography department at University of Georgia for satellite capture of identified locations. The image depicts algal bloom movement captured for following months. The columns show the view for different timeline and rows show different captures of the satellite. The intensity of the green color is the concentration of Chlorophyll A element in the cyanobacteria. In the second row, we understand the bloom

is concentrated on the lower part of the reservoir in the month of June. With the increase in Summer heat, for following months we see the green shade moves to upper region of the reservoir. In the month of August we see major portion of the reservoir is covered under the bloom. By October as the heat reduces the concentration appears to be reducing.

CHAPTER 7

CONCLUSION

In this research, we thus outline the impact of cyanobacteria in worldwide locations. We thus understand why tracking this harmful algal bloom is important. We discuss various approaches utilized for tracking cyanobacteria. We introduce our approach of mining news articles to extract locations affected by cyanobacteria.

In Chapter 2, we study various applications that are dedicated towards cyanobacterial tracking. We discuss their functionality pros and cons. From this study, we understand news articles have so far not been used for cyanobacteria location tracking phenomena. We hence exploit the social sensing aspect by mining news articles.

In Chapter 6, we introduce our approach for location extraction. We discuss how we can extract news articles by integrating Google News API. We further discuss about the unsupervised learning based classification approach to distinguish relevant and irrelevant articles. We further validate our results of this unsupervised learning with results of supervised learning and manual labels. Further we discuss the multiple location keywords issue and we propose our modelling techniques.

We thus conclude by discussing results of each component of the architecture. We also present a list of locations we extracted from news articles having cyanobacterial bloom issue.

7.1 Future Work

We currently are in the process of integrating these locations obtained from news articles with satellite data. The data captured from the satellite provides the intensity of cyanobacterial elements in the bloom based on the images it captures. We hence plan to

integrate the locations from news article with the data obtained from satellite images and develop a map interface on CyanoTracker web application where user will be able to view more technical details of the algae based on the location selection.

Improving on the location modelling parameters is required. Currently we are relying only on two parameters i.e the article headline and keyword frequency. In future we would want to introduce a more semantic approach to perform sentence level classification with context preservation. We also would like to experiment with parameters of AIDA to filter at water body level instead all the location entities.

While this thesis did not explore on semantics such as synonyms, antonyms and inter-relationship between words for location and time extraction, researchers have successfully employed semantics-based approaches for information extraction in several domains such as Bio-Medical, Financial etc [47]. Semantics based approaches typically work by extracting Subject-Predicate-Object triplet from sentences [48]. Specialized parsers such as Standford Parser [49], OpenNLP [50], LinkParser [51] etc are employed for triple extraction. We believe augmenting our framwework with semantics-based approaches will further enhance its capabilities and performance.

REFERENCES

[1] Press Release: Otago Regional Council, "Scoop independant news regional, cyanobacteria found in lake waihola, wednesday, 11 january 2017, 4:33 pm," .

[2] Country News, "Blue-green algae in east loddon water district by country news on march 29, 2017," .

[3] George Francis, "Poisonous australian lake," *Nature International Journal of Science*, vol. 18, 1878.

[4] Wayne W. Carmichael, "The toxins of cyanobacteria," *Scientific American*, vol. 270(1), pp. 78–86, 1994.

[5] Gobler Christopher J. Kramer Benjamin J. Loftin Keith A. Rosen Barry H., Davis Timothy W., "Cyanobacteria of the 2016 lake okeechobee and okeechobee waterway harmful algal bloom," *U.S Geological Survey*, 2017.

[6] CNN Article by Jennifer Gray, "Fish kill in florida: 'heartbreaking images' seen for miles," 2016.

[7] Victor Lawson Michael D. Scott, Lakshmish Ramaswamy, "Cyanotracker: A citizen science project for reporting harmful algal blooms," *Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on*, 2016.

[8] "Google news," 2006.

[9] "Latent dirichlet allocation," *Journal of Machine Learning Research 3 (2003) 993-1022*, 2003.

[10] "Visualizing data using t-sne," *Journal of Machine Learning Research 9 (2008) 2579-2605*, 2008.

[11] Misook Choi Nina Wacholder, Yael Ravin, "Disambiguation of proper names in text," in *Proceeding ANLC '97 Proceedings of the fifth conference on Applied natural language processing*. ACM, 1997, pp. 202–208.

[12] Ilaria Bordino Marc Spaniol Gerhard Weikum Mohamed Amir Yosef, Johannes Hoffart, "Aida: An online tool for accurate disambiguation of named entities in text and tables," in *In: Proceedings of the 37th International Conference on Very Large Databases, VLDB 2011*, 2011, p. 14501453.

[13] KlausBerberich GerhardWeikum JohannesHoffart, Fabian M.Suchanek, "Yago2: A spatially and temporally enhanced knowledge base from wikipedia," *Artificial Intelligence Journal (AIJ)*, vol. 194, pp. 28–61, 2013.

[14] "Cdc-health studies programs-harmful algal bloom(habs)," .

[15] R. O. C. I. O. ARANDA-RODRIGUEZ and ZHIYUN JIN, "evaluation of field test kits to detect microcystins: 2010 study" rapport prpar pour exposure and biomonitoring division health canada, canada," 2011.

[16] "The lake champlain committee," .

[17] "The vermont department of health," .

[18] Vermont departments of Health Lake Champlain Committee and Environmental Conservation, "Cyanobacteria (blue green algae) tracker," 2003.

[19] "Florida fish and wildlife conservation commission," 1999.

[20] U.S Department of Commerce, "National oceanic and atmospheric administration," 1807.

[21] National Oceanic and Atmospheric Administration, "Harmful algal blooms observing system," .

[22] "National aeronautics and space administration (nasa), united states of america," 1958.

[23] "United states geographical survey," 1879.

[24] "United states environmental protection agency, washinton d.c," .

[25] K. Loftin R. P. Stumpf Schaeffer, B. A. and P. J. Werdell, "Cyanobacteria assessment network (cyan)," 2015.

[26] Jeff Hollister Hillary Snook, Shane Bradt, "Cyanobacteria monitoring collaborative - three coordinated projects to locate and understand harmful cyanobacteria, quality assurance program plan, united states environmental protection agency," .

[27] Ken-ichi Ueda Sean McGregor Scott Loarie Nate Agrin, Jessica Kline, "Connect with nature," 2008.

[28] "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] Trond Grenager Jenny Rose Finkel and Christopher Manning, "Proceedings of the 43nd annual meeting of the association for computational linguistics (acl 2005)," in *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. ACL, 2005, pp. 363–370.

[30] "Wikipedia," 2001.

[31] George A. Miller, "Wordnet: A lexical database for english. communications of the acm," vol. 38, pp. 39–41, 1995.

[32] Christiane Fellbaum, "Wordnet: An electronic lexical database. cambridge, ma: Mit press," 1998.

[33] "Princeton university. about wordnet.," 2010.

[34] "Geonames," 2009.

[35] David D. Lewis, "The independence assumption in information retrieval. in proc. of the european conference on machine learning (ecml)," pp. 4–5, 1998.

[36] Shivakumar Vaithyanathan Bo Pang, Lillian Lee, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia*. ACL, 2002, pp. 79–86.

[37] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features. in proc. of the european conference on machine learning (ecml)," p. 137142, 1998.

[38] G. Corrado T. Mikolov, K. Chen and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[39] Geoffrey M. J. Hinton and D. Rumelhart, "Distributed representations," 1986.

[40] N. Yang M. Zhou T. Liu D. Tang, F. Wei and B. Qin, "learning sentiment specific word embedding for twitter sentiment classification,," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. ACL, 2014, vol. 1, p. 15551565.

[41] Y. Zhu J. Lilleberg and Y. Zhang, "support vector machines and word2vec for text classification with semantic features, in cognitive informatics cognitive computing (icci* cc), 2015 ieee 14th international conference on. ieee, 2015," pp. 136–140.

[42] Haixia Liu, "Sentiment analysis of citations using word2vec," 2017.

[43] Google Developers, "The google maps geolocation api, google maps api web services," .

[44] Ekkachai Naenudorn Suphakit Niwattanakul*, Jatsada Singthongchai and Supachanun Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the*

International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong. IMECS, 2013.

[45] Ze-Kai Cheng Feng Qin, Xian-Juan Tang, "Application and research of multi$_l$*abelnaivebayesclassifier*," *vol*.38, *pp*.39 − −41, 2012.

[46] V. Vapnik, "The nature of statistical learning theory. ny: Springer-verlag," 1995.

[47] N. Milic-Frayling J. Leskovec, M. Grobelnik, "learning sub-structures of document semantic graphs for document summarization,," in *In Proceedings of the 7th International Multi-Conference Information Society IS*, 2004, vol. B, pp. 18–25.

[48] Bla Fortuna-Marko Grobelnik Dunja Mladeni Delia Rusu, Lorand Dali, "Triplet extraction from sentences, technical university of cluj-napoca, faculty of automation and computer science g. bariiu 26-28, 400027 cluj-napoca, romania department of knowledge technologies, joef stefan institute jamova 39, 1000 ljubljana, slovenia," .

[49] "Stanford parser webpage, http://nlp.stanford.edu/software/lex-parser.shtml," .

[50] "Welcome to apache opennlp, https://opennlp.apache.org/," .

[51] "Link parser webpage, http://www.link.cs.cmu.edu/link/," .