Sentiment Analysis of Canadian News Headlines through Cloud-Based Machine Learning

Thesis Paper

By: Dipeeka Luitel

April 14, 2020

Supervisor: Anthony Pagnotta

Algoma University

Department of Mathematics and Computer Science

Abstract

Machine learning is a growing field with many use cases, including text classification. In the current social climate of media organisations using fear mongering news headlines to generate revenue, we seek news content that is less desolate. This paper will examine cloud-based machine learning to recognize newspaper articles that have an optimistic topic or uplifting content. The research question is as follows: is AWS cloud-based machine learning algorithm effective enough for identifying the sentiment of CBC news headlines. The effectiveness was measured by the certainty probability of the machine learning algorithm. The scoring and evaluations of the results produced by the algorithm was deemed insignificant, however there were several discoveries made on how to improve upon this research.

Acknowledgements

I would like to sincerely thank my thesis advisor, Professor Anthony Pagnotta, for his invaluable guidance, continuous support, and for inspiring me to pursue the topic of cloud computing for which I am very grateful. I would also like to thank Algoma University and the Department Chair for the Computer Science program, Dr. Simon Xu, for providing me with the tools to complete this thesis.

Special thanks to my mom for the endless supply of kindness and tea.

Table of Contents:

1. Introduction

Introduction of Thesis

As technology has evolved and encompassed new areas of day-to-day life, computers have impacted virtually every aspect of how people interact with each other and how they interact with the world around them. Information technologies such as the prevalence of cloud-based websites and applications have trivialized the classical information availability problems including scarcity of knowledge to improve the lives of many around the world. The ease and accessibility of communication through technological leaps in our infrastructure have enabled novel and faster methods of transferring information both at a local and a global scale. News and media is a form of communication which, while has benefited from these advances in technology, has also led to a darker undercurrent of apathy towards the social, economic, and political issues around us. The objective of many modern media companies has shifted from being a reliable source of information to society to using anxiety inducing breaking news headlines to gain the most clicks or have more views. While this can make for higher advertising profits for media corporations, it can be easy for the everyday person to become overwhelmed with pessimism and unhappiness. The purpose of this thesis is to counteract this trend by applying cloud-based machine learning natural language recognition to decipher and attempt to understand the amount of happiness contained in a newspaper article, where happiness is defined as an optimistic and uplifting topic.

To understand machine learning, the term Artificial Intelligence must first be defined. "Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial

creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons)" [33]. The term machine learning can be defined as: "... a subset of the larger field of artificial intelligence (AI) that focuses on teaching computers how to learn without the need to be programmed for specific tasks …" [17]. Further exploring the definition of machine learning, it is a discipline that aims to solve the interrelated questions of: how can a computer system be constructed which improves automatically with experience, and what are the fundamental statistical-computational-information—theoretic laws that govern all learning systems, including computers, humans, and organizations [24].

An overview of the machine learning process can be observed in the following diagram:

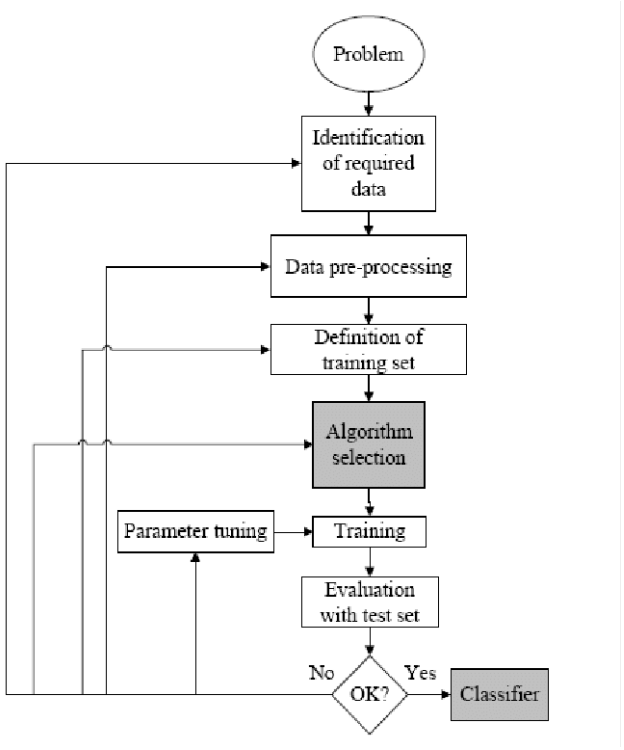


Figure 1: Process of Supervised Machine Learning
Source: Adapted from [3]

The process of machine learning has numerous steps, but the first step is perhaps the most important which is data preparation. “Data preparation typically involves these tasks: Select a sample subset of data**.** Make and track assumptions about the data to select attributes germane to the problem you want the algorithm to train for or solve. For example, filter or focus on types of product or customer data and eliminate data about where a product was manufactured. Merge or join data sets to aggregate records. Merging simplifies the data and makes it easier to manage. For example, if there is a customer data set and a customer purchases data set, they could be condensed into a new, simpler, attribute for spending for the product. Format and sort the data for modeling**.** Choose the format: flat file or relational database for example. Certain algorithms may require data to be sorted in a specific way. For example, fields for customers may be grouped by where the customer purchased or where they live. These textual, location fields may need to be given numbers and sorted numerically. Clean the data by removing or replacing any blank or missing values. There are statistical analysis tools that can help inspect the data for errors and deviations. The goal is to ensure that data is exact, complete and relevant. Normalize the data or adjust values that are measured on different scales to a common scale. For example, one data set may score numerically and another by a percentage. To compare the data, the values must be normalized to a common scale” [18].

After the first two steps of data collection and data preparation, machine learning is performed in the following order with the first step being to select a model for the algorithm, whether this be predefined, which is the method chosen for this research, or a uniquely created algorithm. The next step is the training of the model and “is the process of analyzing input data by using the parameters of a predefined model. From this analysis, the model learns the patterns,

and saves them in the form of a trained model" [38]. This is followed by the scoring of the model, this is "… also called prediction, and is the process of generating values based on a trained machine learning model, given some new input data. The values or scores that are created can represent predictions of future values, but they might also represent a likely category or outcome. The meaning of the score depends on the type of data you provide, and the type of model that you created" [27]. The final step is the evaluation of the model which "…is performed after training is complete, to measure the accuracy of the predictions and assess model fit" [26].

This thesis aims to use natural language processing to examine news media headlines. "Natural language processing (NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language. NLP research has evolved from the era of punch cards and batch processing, in which the analysis of a sentence could take up to 7 minutes, to the era of Google and the likes of it, in which millions of web pages can be processed in less than a second. NLP enables computers to perform a wide range of natural language related tasks at all levels, ranging from parsing and part-of-speech (POS) tagging, to machine translation and dialogue systems" [37]. "Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important" [42]. Although there have not been many investigations on my particular topic, there is similar research on non-cloud-based machine learning and overall happiness using twitter data which can be used as valuable research on this particular topic. The outcome of this thesis is to utilize a cloud-based machine learning system to

identify and rank the happiness of a particular news article, and thus enable readers to choose articles which will provide them with a more positive outlook in their day to day life.

A brief overview of the literature review includes several peer-reviewed academic papers, including *A Machine Learning Approach to Predict Happiness Based on Sentiment Analysis of Twitter Data* [21], that was used as a basis for the research on my topic. There will also be several textbooks used as a source of reference for the topic ranging from a general introduction to machine learning to cloud-based machine learning, and also specifically using the Amazon AWS platform for utilizing machine learning in the cloud. These textbooks provide a strong knowledge foundation base for research on the topic.

The materials that were used for this thesis is a 2017 MacBook Pro, Amazon AWS SageMaker via a supervised learning algorithm, datasets from an the open source website *Kaggle* (*www.kaggle.com*) and dataset purchased from the *Linguistic Data Consortium*, and articles from the newspaper organization CBC. The SDLC that was used throughout this project is the iterative and incremental development method. Supervised learning can be defined such that "… the training set includes labels so the algorithm knows the correct label given a set of attributes. For example, the attributes could be the color and weight of a fish where the label is the type of fish. Eventually the model learns how to assign the correct or most probable label. A typical supervised learning task is classification, which is the task of assigning inputs such as text or images to one of several predefined categories" [6].

It should be noted that “the algorithms used in this category typically consist of k-nearest neighbors, linear regression, logistic regression, support vector machines, and neural networks” [6]. Even within the supervised learning subset, there are numerous types of algorithms that must be considered when choosing which will work best for the desired research outcome. The BlazingText algorithm, which was ultimately chosen for the research experiment, uses the Word2vec algorithm which “... maps words to high-quality distributed vectors. The resulting vector representation of a word is called a word embedding. Words that are semantically similar correspond to vectors that are close together. That way, word embeddings capture the semantic relationships between words.” [7].

Two of the main types include the classification algorithm and the regression algorithm. “Classification algorithms are used when the desired output is a discrete label. In other words, they’re helpful when the answer to your question about your business falls under a finite set of possible outcomes. Many use cases, such as determining whether an email is spam or not, have only two possible outcomes. This is called binary classification” [29]. Regression on the other “... is useful for predicting outputs that are continuous. That means the answer to your question is represented by a quantity that can be flexibly determined based on the inputs of the model rather than being confined to a set of possible labels. Regression problems with time-ordered inputs are called time-series forecasting problems, like ARIMA forecasting, which allows data scientists to explain seasonal patterns in sales, evaluate the impact of new marketing campaigns, and more” [29].

1.1 Review of the Literature:

Although the term machine learning has been misappropriated as a marketing buzzword in recent years, the use of cloud-based machine learning for solving complex computer science problems is something that is still a relatively new field of study. There have been a variety of platforms that allow for clients to perform machine learning as a service (MLaaS) including Amazon Machine Learning, Microsoft Azure Machine Learning, and Google Cloud Machine Learning [16]. To fully understand the functionality of MLaaS we first need to explore *as a Service*, often denoted as XaaS. The textbook *Cloud Computing for Science and Engineering* outlines this definition as a company, such as Amazon Web Services, providing a system, such as machine learning or perhaps another service such as software like natural language processing via AWS Comprehend, to a large customer base [16].

The Cloud Computing for Science and Engineering textbook mentioned above provided thorough research on each of the aforementioned cloud platforms and the respective advantage that each platform holds. In brief summary, it stated that Amazon's machine learning was performed through creating a predictive model based on training data provided by the customer [16]. Amazon has a variety of artificial intelligence services such as Comprehend (used for natural language processing), Lex (uses voice input into applications), and also machine learning services such as SageMaker [16].

Microsoft Azure Machine Learning on the other hand is based on a drag-and-drop component composition model, that creates a solution from connecting tools together into a workflow graph. Amazon is seen as the market leader in providing computing, storage, and

platform services [16]. Azure services machine learning as an on-demand service that does not require the customer to deploy and manage a virtual machine for the service [16]. Azure is seen as the second biggest player in cloud infrastructure, but due to the nature of their drag-and-drop machine learning platform was ultimately rejected for this thesis project. Lastly, while Google Cloud does provide machine learning services, this platform offered the smallest range of support and features. Based on these qualities listed for each of these platforms, it became clear that Amazon as a service provider best fulfilled the needs of this thesis requirement. This textbook was thus a very valuable book for both the initial exploration conducted on the topic of various cloud platforms and was utilized as a touchpoint throughout the research investigation.

There are several sources that look at the use of machine learning, whether they be cloud-based or not, to observe the happiness factors of a specific topic. One of these sources, *Meaningless comparisons lead to false optimism in medical machine learning* [31] by authors Orianna DeMasi, Konrad Kording, and Benjamin Recht, has the purpose of determining how healthcare is being impacted by the use of sweeping algorithms for making assessments of patient with a baseline that may not be significant to the patient. The findings of this paper determined that machine learning could be poorly trained and lack the ability in being able to find the realizations of data from the actual content.

The paper *A Machine Learning Approach to Predict Happiness Based on Sentiment Analysis of Twitter Data* [21] by Satyabrata Aich, Ki-Won Choi, and Hee-Cheol Kim, looks at using the keyword of ‘happiness’ as well as ‘sadness’ to guess the sentiments of Korean twitter users. Utilizing a single keyword or phrase would provide a simplified method of approach for

providing training data materials for machine learning. *An Exploratory Study on Machine Learning Model Stores* [25] by Minke Xiu, Zhen Ming Jiang, and Bram Adams examines the differences between various machine learning models including AWS marketplace, Google Play, and Apple's App Store.

There are also several textbooks used as a source of research for the topic ranging from a general introduction to machine learning to cloud-based machine learning, and also specifically using the Amazon AWS platform for utilizing machine learning in the cloud. These textbooks provide a strong foundation of knowledge for the research on this topic. Some of these titles include: *Machine Learning Pocket Reference: Working With Structured Data in Python* [23], *Developing Information Systems - Practical guidance for IT professionals* [15], *Machine Learning for dummies* [20], *Machine learning with AWS* [19], and *Pragmatic AI* [30].

The textbook *Artificial Intelligence: A Modern Approach* was used to provide a deeper fundamental understanding on the basics of artificial intelligence (AI), it was also used to examine the history of AI and how it evolved to encompass machine learning as we know it to be today with various algorithm searching methods [36].

1.2 Objective, Aim and Research Question(s):

This thesis paper will research cloud-based machine learning to recognize newspaper articles that have an optimistic topic or uplifting content. The research question is as follows: is AWS cloud-based machine learning algorithm effective enough for identifying the sentiment of CBC news headlines. The effectiveness was measured by the certainty probability of the

machine learning algorithm. This thesis defined having that this minimum accuracy will be of 75%, as this is the standard requirement that would deem the results as a success. The aim was to question if the AWS cloud-based machine learning algorithm is efficient for identifying the sentiment of CBC news headlines.

1.3 Rationale:

The decision that led to this topic being chosen for research and evaluation was simply due to the lack of optimism when opening a newspaper or checking a news app for keeping up-to-date with the daily news. It can become increasingly difficult to continue being invested in important topics when apathy clouds the judgement. The constant bombardment of negative, attention grabbing headlines is not something new in the modern age of technology. As technology brings people in society closer together through various social media platforms, the availability of millions of databases online, and infinite knowledge on the internet, it is also being used to keep members of society feeling alienated and alone with the constant assault of negativity throughout the world as portrayed by news media organizations. The use of fear and invoking panic is a very powerful tool to get consumers to purchase products so they feel comfortable and safe, whether that be unnecessary material goods or viewership revenue gained from fear mongering headlines.

This persuasive tactic can be observed with the COVID-19 pandemic, currently ongoing at the time of research for this thesis paper. The pandemic has led to people rushing into grocery stores en-masse to hoard food and toiletries even though these stores remain open and are deemed as essential services. An interesting example of this can be observed with many

countries having toilet paper shortages due to consumer hoarding. There should be no rational cause for needing to hoard materialistic goods to feel safe, but the medical pandemic has in turn created a fear pandemic which is being propagated by news media organizations. This is beneficial to them because having the general public live in fear ultimately raises their viewership and as a result their revenue and profit. Therefore, it is important to have a balance of positivity to counteract the constant negativity, because it allows the everyday user to maintain a degree of happiness without becoming indifferent or irrational to the pessimistic events and news that occur worldwide. This rationale is the reason this topic was selected for this thesis research.

1.4 Scope:

The scope of this thesis will cover major news headlines from the CBC news media organization from the previous years. The potential areas of limitations for this research includes topics out of the Canadian news media perspective and priorities. Additionally, this thesis will not be able to examine a large quantity of data in the media coverage of years (this limitation is simply due to the timeline of the project which does not allow sufficient time for vast volumes of collection and creation in the dataset) which may limit the full quantitative overview of how the sentiment used to describe headlines of media companies has changed over the duration of the scope.

1. 5 Timetable:

Item	Date
Thesis Proposal Submission	Thursday November 28, 2019
Proposal Presentation Practice	Tuesday December 3, 2019
Public Proposal Presentation	Friday December 6, 2019
Submission of Written Report	Tuesday March 31, 2020
Final Presentation Practice	Thursday April 2, 2020
Public Thesis Presentation	Saturday April 4, 2020
Submission of Final Version of Written Report	Tuesday April 14, 2020

2. Methods

2.1 Methods

The process of conducting the scientific experiment for this thesis was broken up into four major iterative sections along with an initial planning and a deployment stage. Prior to the start of the research, the initial environment needs to be set up. This involves getting proper IAM roles and creating an Amazon Simple Storage Service (S3) bucket to store all data. "Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. This means customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides easy-to-use management features so you can organize your data and configure finely-tuned access controls to meet your specific business, organizational, and compliance requirements" [8]. The use of S3 buckets in this project was to store the input training data sets and any additional data, hold output files created by the machine learning system , and hold any extra materials that were necessary for the research and experiment.

An IAM role is a more complex topic and it should be noted that for this experiment to be properly functional there must be full privileges on IAM. "An IAM *role* is an IAM identity that you can create in your account that has specific permissions. An IAM role is similar to an IAM user, in that it is an AWS identity with permission policies that determine what the identity can and cannot do in AWS. However, instead of being uniquely associated with one person, a role is

intended to be assumable by anyone who needs it. Also, a role does not have standard long-term credentials such as a password or access keys associated with it. Instead, when you assume a role, it provides you with temporary security credentials for your role session. You can use roles to delegate access to users, applications, or services that don't normally have access to your AWS resources. For example, you might want to grant users in your AWS account access to resources they don't usually have or grant users in one AWS account access to resources in another account. Or you might want to allow a mobile app to use AWS resources, but not want to embed AWS keys within the app (where they can be difficult to rotate and where users can potentially extract them). Sometimes you want to give AWS access to users who already have identities defined outside of AWS, such as in your corporate directory. Or, you might want to grant access to your account to third parties so that they can perform an audit on your resources. For these scenarios, you can delegate access to AWS resources using an *IAM role*" [14]. To use the functionalities of Amazon SageMaker, there must first be a SageMaker notebook instance created. A notebook instance is "fully managed ML compute instance running the Jupyter Notebook App. Amazon SageMaker manages creating the instance and related resources. Use Jupyter notebooks in your notebook instance to prepare and process data, write code to train models, deploy models to Amazon SageMaker hosting, and test or validate your models" [9].

The first section involved the creation of a machine learning (ML) algorithm. The environment in which the experiment was performed is Amazon Web Services, which is highly supportive of Python as its language of choice for machine learning. In addition to the standard Python libraries, the machine learning libraries were used in order to better facilitate efficiencies of vector mapping for sentiment analysis. There was also code written in the Java language

which was used for data preparation and data cleaning. Each algorithm on AWS has its own requirements for the input file that is accepted, so the data preparation must take this into account. There were in total three classes created which were used in incremental steps to ensure the data sample was prepared with the proper input format for the BlazingText requirements. The first Java class had the functionality to extract article headlines from larger datasets, this ensured that no unnecessary values were input for the supervised learning. The second Java class had the functionality to remove any unwanted characters from the newly created .txt file to reduce any noise that the file may have had. The last Java class had the functionality to format each headline in the proper manner with a label index "__label__" prior to the start of the line and a delimiter of " ." to end the headline. This ensured that all headlines were uniform and the dataset .txt file would be properly accepted by the BlazingText algorithm.

The second iteration will include several rounds of training of various data sets to ensure the algorithm produced is reliable. Amazon Web Services offers several of its own unique machine learning algorithms in its repertoire, and after some initial testing found that this algorithm best suits the needs of the research and thus this project aimed to use the XGBoost algorithm as part of its overall arching algorithm, however during the actual experiment it was discovered that the BlazingText algorithm was best suited for the needs of this project.

The third step will include rigorous testing of the algorithm and gathering of data on the outcomes of the test results, both through the use of unit testing and also integration and acceptance testing and optimization. The experiment had insufficient time to request the

approval for utilization volunteers in performing user testing of this experiment by the Algoma University Research Ethics Board.

Lastly, the data collected was evaluated to gain a clear understanding of how the algorithm performed and see if any necessary steps need to be performed to modify the algorithm. Should any revision be required, the aforementioned steps of life cycle were iterated through to ensure any errors are corrected and the results are filtered for possible errors.

After finishing the iterations, in the deployment step there was testing performed on the algorithm – both by the researcher through unit testing, interaction testing, as well as acceptance testing. Had the user testing b include having five individuals deploy the algorithm and interpret the results and share their thoughts afterwards. The data was gathered, and all results and evidence was used to reach a conclusion for the thesis topic.

2.2 Materials

The materials for this project are mostly software based as this research paper deals with cloud computing, therefore the specifications of the computer are less vital to the overall development and procedures for the experiments performed. In this specific paper, the computer used is the 2017 MacBook Pro 13-inch model and the specs of the computer used include: "**Retina display -** 13.3-inch IPS technology 2560x1600 pixels. P**rocessor –** Core i5 2.3 (Kaby Lake 15W) **Storage**- 128GB SSD**,** Configurable to 256GB, 512GB, 1TB or 2TB SSD. **Memory -** 8GB configurable to 16GB of memory. **Graphics -** Intel Iris Plus Graphics 640. **Charging and Expansion -** Two Thunderbolt 3 (USB-C) ports and 3.5mm headset. **Wireless - Wi-Fi** 802.11ac

Wi-Fi wireless networking**,** IEEE 802.11a/b/g/n compatible and **Bluetooth** 5.0 wireless technology. **Size and Weight - Height:** 1.49 cm (0.59 inch)**, Width:** 30.41 cm (11.97 inches)**, Depth:** 21.24 cm (8.36 inches)**, Weight:** 1.37 kg (3.02 pounds). **Operating System - macOS** Catalina" [2].

Additionally, the software used for this is SageMaker, a machine learning as a service (MLaaS), on Amazon Web Service. "Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly. SageMaker removes the heavy lifting from each step of the machine learning process to make it easier to develop high quality model" [4]. The choice behind using Amazon Web Service can be is due to it being one of the largest companies that offer of cloud computing and is very comprehensive in its support of MLaaS and cloud-based storage for containing all data for the entirety of the project, and other XaaS that may be used to evaluate the results of this research project. The language used in the SageMaker instance was Python 3 on Jupyter, the reasoning behind this language choice was due to it being a prominent language in the field of both machine learning and Amazon Web Services which results in a large availability of library support for the creation of the algorithm. The Java language was also used for data preparation and data cleaning. The version used was Java 8 Update 201 and the integrated development environment used alongside Java was Eclipse IDE 2019-12.

The configuration settings used when creating the SageMaker notebook instance is as follows: the *Notebook instance type* is "ml.t2.medium", the *Elastic Inference* is set to "none", *IAM role* was set to Create a new role with access for S3 set to "Any S3 bucket" attached to the

account. The optional settings: *Network* was set to "No VPC", *Git repositories* was set to "None", and *Tags* did not have any keys or values added. After all the proper modifications were made to the settings, the notebook instance was created.

A component to the algorithm that was used is the XGBoost (eXtreme Gradient Boosting), which is an implementation of the gradient boosted trees algorithm [43]. "Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler, weaker models" [43]. Therefore, this makes "… XGBoost a solid choice for problems in regression, classification (binary and multiclass), and ranking" [43] "This current release of the XGBoost algorithm makes upgrades from the open source XGBoost code base easy to install and use in Amazon SageMaker. Customers can use this release of the XGBoost algorithm either as an Amazon SageMaker built-in algorithm, as with the previous 0.72-based version, or as a framework to run training scripts in their local environments as they would typically do, for example, with a TensorFlow deep learning framework. This implementation has a smaller memory footprint, better logging, improved hyperparameter validation, and an expanded set of metrics than the original 0.72-based version. It also provides an XGBoost estimator that executes a training script in a managed XGBoost environment. The current release of Amazon SageMaker XGBoost is based on version 0.90 and will be upgradeable to future releases" [43]. During experimentation various algorithms were experimented with however ultimately the BlazingText Algorithm was chosen. BlazingText uses "… downstream natural language processing (NLP) tasks like sentiment analysis, named entity recognition, and machine translation require the text data to be converted into real-valued vectors" [7]. "BlazingText's highly optimized implementation of the Word2Vec

algorithm, for learning these vectors from several hundreds of gigabytes of text documents. The resulting vectors capture the rich meaning and context that we recognize when we read a word. BlazingText being more than 20x faster than other popular alternatives like fastText and Gensim, enables [users] to train these vectors on their own datasets containing billions of words using GPUs and multiple CPU machines, hence reducing the training time from days to minutes" [35].

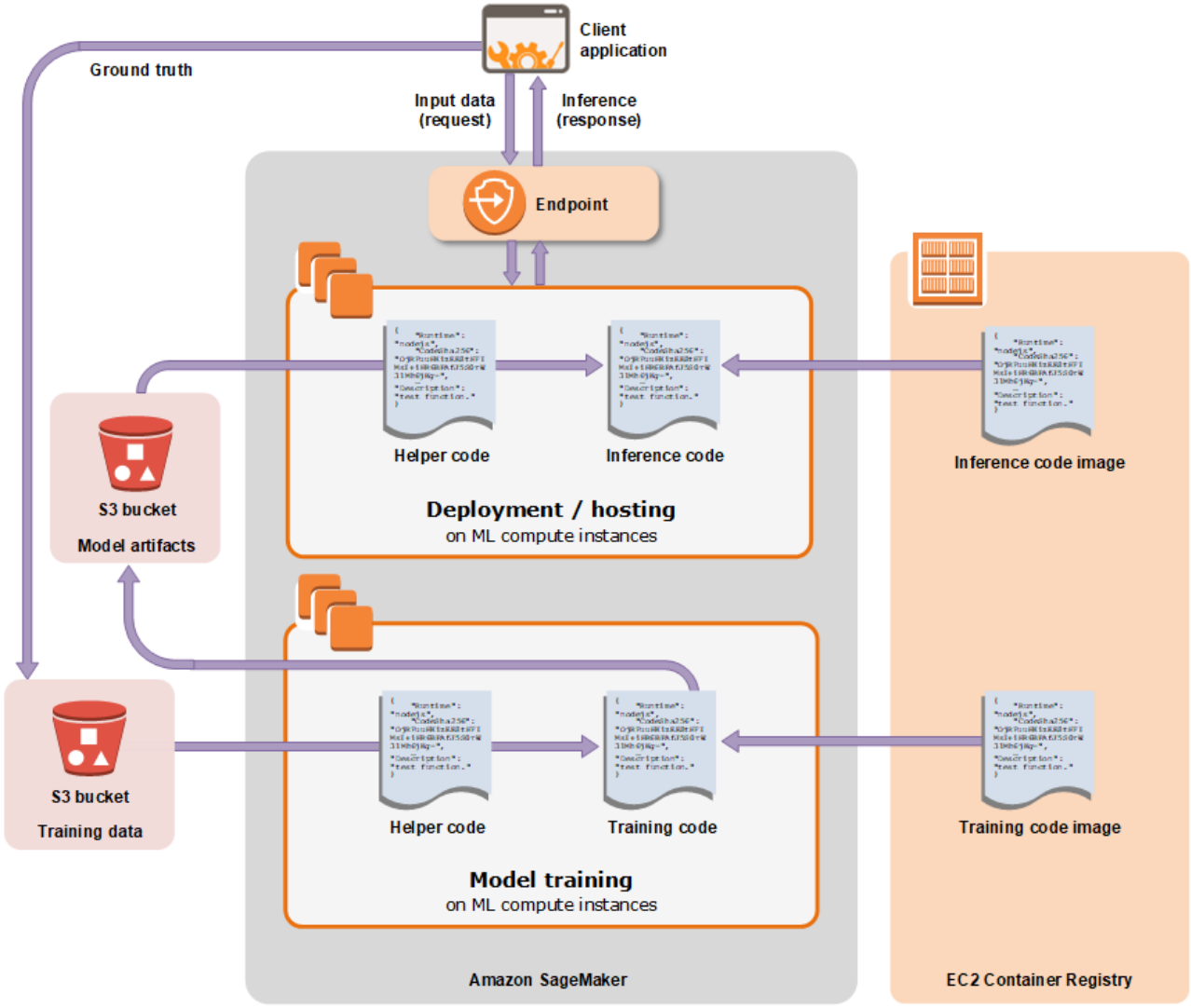


Figure 2: AWS SageMaker
Source: Adapted from [39]

The three datasets that were used in this research include: A Million News Headlines [1], "The AQUAINT Corpus of English News Text" which was purchased from the *Linguistic Data Consortium* (LDC) database [12], and the final dataset was a compilation of twenty different articles from the CBC news media website. The headlines of the articles used in the machine learning are as follows:

- Article 1: We will get through this: Oilpatch hunkers down amid price plunge virus fears
- Article 2: South Korea taking 'unprecedented' steps as Italy, Iran also struggle to contain COVID-19
- Article 3: COVID-19 in Quebec: Province up to 2,840 confirmed cases, but premier sees encouraging signs
- Article 4: British Columbians stranded abroad feel left in the dark by government
- Article 5: Olympics postponement raises questions, throwing athletes' scheduling into disarray
- Article 6: Quebec biotech firm produces a potential COVID-19 vaccine
- Article 7: Coronavirus: WHO calls COVID-19 outbreak a pandemic as Italy orders most stores to close
- Article 8: Quebec's first specialized COVID-19 clinic opens in Montreal
- Article 9: BC Hydro says customers impacted by COVID-19 can ask for help with bill payments
- Article 10: N.B. COVID-19 roundup: Province braces for 'next big wave' of coronavirus
- Article 11: Number of COVID-19 cases surpasses 100,000 worldwide
- Article 12: University of Alberta abandons letter grades, cancels most exams amid pandemic

- Article 13: Alberta radiologists 'bewildered and demoralized' as province cancels contracts amid COVID-19 pandemic
- Article 14: B.C. premier vows province will meet the challenge of COVID-19
- Article 15: 'We can't get home': Stuck in limbo abroad, these Winnipeggers wait
- Article 16: P.E.I. Premier Dennis King declares 'public health emergency' on COVID-19
- Article 17: Canadians trapped in Morocco by COVID-19 restrictions to be evacuated this weekend: Trudeau
- Article 18: 'Help is on the way' for renters during coronavirus crisis, says B.C. housing minister
- Article 19: Get us out: Canadians still stranded abroad wait to hear if Ottawa will help them
- Article 20: Quebec announces first death from COVID-19 confirmed cases up to 94

Screenshots of each individual article mentioned above can be viewed in the appendix.

2.3 Description of Software Development Life Cycle Model

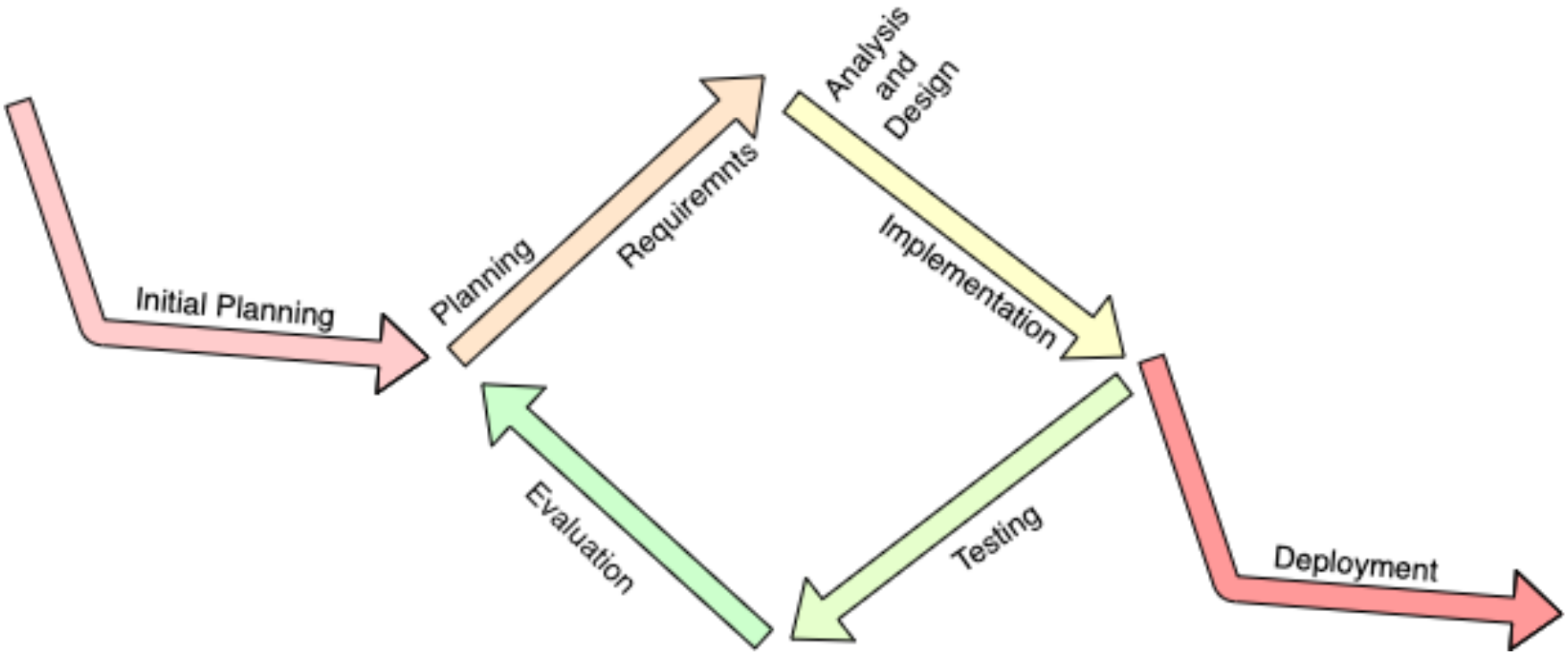


Figure 3: Iterative and Incremental Software Development Life Cycle, based on the figure from *Developing Information Systems - Practical guidance for IT professionals* [13].

The software development life cycle used in this research paper can be seen in Figure 2 as the Interactive and Incremental design. The reason that this SDLC was chosen for the research project was because it allowed for more flexibility in any iterations needed for the life cycle unlike the traditional Waterfall Model, and it also did not carry the burden of issues that arise when using Agile.

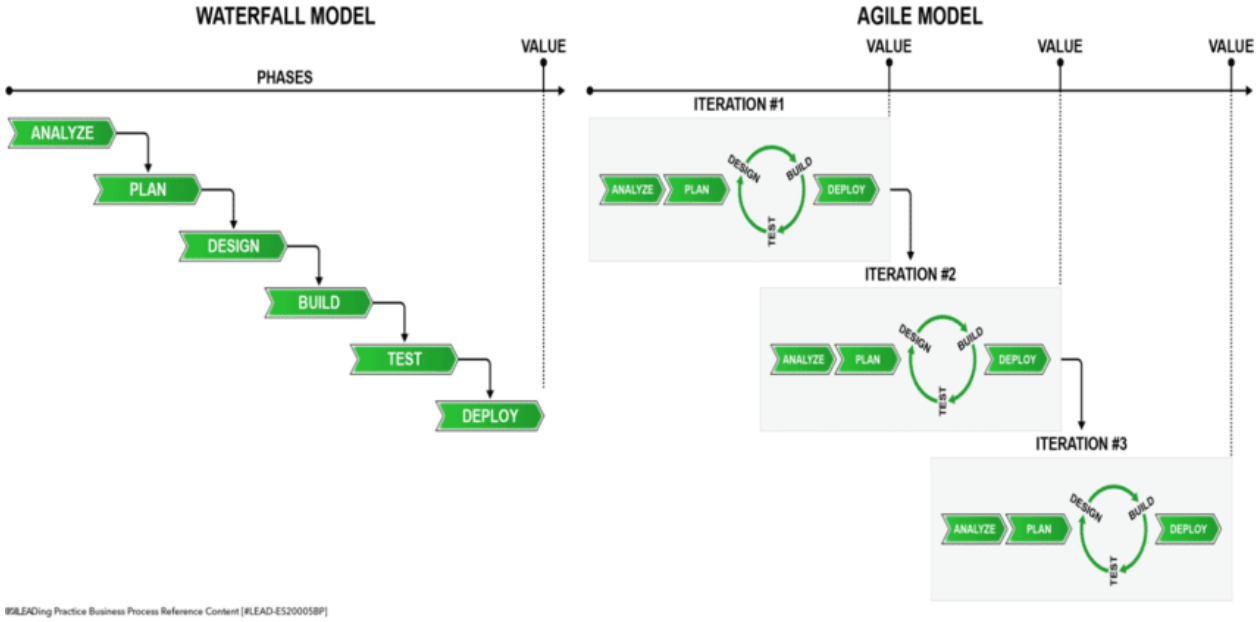


Figure 4: Traditional Waterfall versus Agile Model [28].

The Waterfall Model pays attention to planning the architecture and structure of the software system in detail [28] This may be beneficial but since the Waterfall Method is a highly structured, linear model this leaves little room for going back to correcting any errors after leaving a stage that has been completed [28]. "Small projects are almost always suitable for an Agile approach and almost never for a Waterfall approach. Both the Waterfall Model and the Agile Methodologies have their issues when tackling medium-sized projects. The demanding Waterfall Model could add too much overhead to a fairly easy project, while a flexible Agile

Methodology could be too easy going for the same project" [41] In this particular case, the other of Agile including Daily Scrum and Sprints would be negated with a one-person team and so Agile was ultimately not a good choice of SDLC to use in the implementation of the project.

The path that was used for this lifecycle is broken down to the following stages:

- The initial planning of this project involves development of the research, including the scope, the data objective, and the data input/output.
- The planning stage comprises writing python code that was used to create a supervised machine learning algorithm and the objective was to intake news articles into the program and evaluate the quality of how optimistic the content of the article is. The data input requirement was a .txt file of media headlines and article contents which revolve around the topic of novel coronavirus or COVID-19 ranging from the time span of February 2020 to March 2020 from the Canadian Broadcast Corporation (CBC) online news website to train and test the algorithm.
- The design of the algorithm ensures that once the data set is obtained, there must be data cleaning performed (and there must be a training set and a test set created), as well as training and scoring of the machine learning model before the model is evaluated.
- The process of training the data set is then performed to ensure the model was able to accurately and reliably test the data. The data output was the evaluation performed via the algorithm which generated a numerical breakdown of the diction and language used in the article correlates to the optimism of the content. After any further iterations are completed as required, the algorithm was deployed for the next phase – testing of the product.

2.4 Software Test Plan

The software test plan for the testing of the ML algorithm used includes a variety of testing to ensure all aspects of the completed project are sufficient and fulfil their objectives. The forms of testing included:

- Unit Testing to ensure that individual components of the code correctly perform the required task, modules to be tested include functions that opens and parses text files with the potential risk being that the text files cannot be opened due to them being corrupted. The tentative testing period is March 14 – 17th, 2020.
- Integration testing to ensure the components are cohesive in working together to accomplish the machine learning tasks, this would test the whole system with potential risk involved being that some functionality being depreciated for certain methods. The tentative testing period is March 18 – 20th, 2020.
- Acceptance testing to ensure that all of the original requirements of the program have been met, and the potential risks would include some requirements not being met by the system. The tentative testing period is March 20 – 21st, 2020.

User testing would have been beneficial to gain a different view of how the product is used by an external demographic to the researchers. The area to be tested would have been the whole system which will allow a deeper understanding of how an external demographic interacts with the system. Potential risks included the potential to perhaps reveal errors or oversights in the system that the research failed to take into account. This form of testing was unfortunately not included in this research due to limitations in the timeline of this project due to the long duration of time necessary to acquire the proper approvals from the Algoma

University Research Ethics Board "… to create a research environment in which the University's responsibilities towards Human Participants involved in research are discharged in accordance with the highest ethical standards; to promote awareness and understanding of such standards among members or associated members of the University…" [11].

The software testing methods described above can all be summarized in the figure below:

Software Testing Method	Test Objective(s)	Module(s) to be Tested	Potential Risks Involved	Testing Period
Unit Testing	Testing of individual software components that make up system	Function that opens and parses text files	Text files can be corrupted. Text files cannot be used by the algorithm due to special characters	03/14/2020 – 03/17/2020
Integration Testing	Components tested when integrated together to perform specific tasks and activities	The whole system	Function(s) to be integrated may not be the most up-to-date versions	03/18/2020 – 03/20/2020
Acceptance Testing	Making sure that all the system requirements have been met	The whole system	Some key requirements are not being taken into account	03/20/2020 – 03/21/2020

Table 1: based on the figure from *Updating your Methods section* [22].

3. Results

3.1 Introduction

Through the various iterations of this project, a large quantity of data was produced. However, this paper seeks to mainly examine the output of results created during the last and final iteration. As stated above in the previous section, the main purpose of the research performed is to answer the research question as follows: is AWS cloud-based machine learning algorithm effective for identifying the sentiment of CBC news articles, with effectiveness being measured by the certainty probability of the machine learning algorithm. The aim, on the other hand, was to question if the AWS cloud-based machine learning algorithm is efficient for identifying the sentiment of CBC news articles. The efficiency of the algorithm was measured in a duration of seconds. The results also reviewed the proposed forms of testing for this study which include the following: unit testing, integration testing, acceptance testing, and user testing. Due to limitations of this study, user testing was unable to be performed as a method of evaluation.

The purpose of this research was to develop a ML model that can detect the sentiment of the news article headlines that are being analyzed with an accuracy of a certain percentage. This thesis defined that by having an accuracy of 75%, this would deem the results as being successful. Anything lower than that would be considered a failure on a gradient. In a brief overview, the data and certainty percentage in the accuracy of output gathered from the supervised machine learning state shows that overall the system had room for significant improvement. One cause of this can be explained due to data cleaning for this specific use case

of the BlazingText algorithm. Through the various iterations of training and testing the data, there may not be sufficient data volume to make an impact on proper training of the algorithm combined with the algorithm being inefficient fundamentally in the approach it took to tackle the problem. The action of cleaning the headlines may have been a threat to accuracy in the certainty probability, which is a result of the algorithm intake of data that does not take into consideration the special characters that were removed for data cleaning. The recommendations on how to improve results for further testing will be discussed in Chapter 4 of this paper.

3.2 Main Findings

The following screen images below display each step of software testing performed in the Jupyter notebook to reach the results gathered. The full code can be viewed in the appendix.

In [1]:
```
import sagemaker
import boto3
import pandas as pd
from sagemaker import get_execution_role
import json

sess = sagemaker.Session()

role = get_execution_role()
print(role)

region = boto3.Session().region_name
bucket = 'dluitel'
print(bucket)
prefix = 'sagemaker/blazingtext'
bucket_path = 'https://s3-{}.amazonaws.com/{}'.format(region,bucket)
print(bucket_path)
```

```
arn:aws:iam::524541277508:role/service-role/AmazonSageMaker-ExecutionRole-20200329T185578
dluitel
https://s3-us-east-1.amazonaws.com/dluitel
```

Figure 5: Cell [1]

In [2]:
```
import io
import boto3
import random

def data_split(FILE_DATA, FILE_TRAIN, FILE_VALIDATION, FILE_TEST, PERCENT_TRAIN, PERCENT_VALIDATION, PERCENT_TEST):
    data = [l for l in open(FILE_DATA, 'r')]
    train_file = open(FILE_TRAIN, 'w')
    valid_file = open(FILE_VALIDATION, 'w')
    tests_file = open(FILE_TEST, 'w')

    num_of_data = len(data)
    num_train = int((PERCENT_TRAIN/100.0)*num_of_data)
    num_valid = int((PERCENT_VALIDATION/100.0)*num_of_data)
    num_tests = int((PERCENT_TEST/100.0)*num_of_data)

    data_fractions = [num_train, num_valid, num_tests]
    split_data = [[],[],[]]

    rand_data_ind = 0

    for split_ind, fraction in enumerate(data_fractions):
        for i in range(fraction):
            rand_data_ind = random.randint(0, len(data)-1)
            split_data[split_ind].append(data[rand_data_ind])
            data.pop(rand_data_ind)

    for l in split_data[0]:
        train_file.write(l)

    for l in split_data[1]:
        valid_file.write(l)

    for l in split_data[2]:
        tests_file.write(l)

    train_file.close()
    valid_file.close()
    tests_file.close()

def write_to_s3(fobj, bucket, key):
    return boto3.Session(region_name=region).resource('s3').Bucket(bucket).Object(key).upload_fileobj(fobj)

def upload_to_s3(bucket, channel, filename):
    fobj=open(filename, 'rb')
    key = prefix+'/'+channel
    url = 's3://{}/{}/{}'.format(bucket, key, filename)
    print('Writing to {}'.format(url))
    write_to_s3(fobj, bucket, key)
```

Figure 6: Cell [2]

In [3]:
```
import urllib.request

FILE_DATA = 'uci-headlines'

#split the downloaded data into train/test/validation files
FILE_TRAIN = 'uci-headlines.train'
FILE_VALIDATION = 'uci-headlines.validation'
FILE_TEST = 'uci-headlines.test'

PERCENT_TRAIN = 70
PERCENT_VALIDATION = 15
PERCENT_TEST = 15

data_split(FILE_DATA, FILE_TRAIN, FILE_VALIDATION, FILE_TEST, PERCENT_TRAIN, PERCENT_VALIDATION, PERCENT_TEST)
```

Figure 7: Cell [3]

In [4]:
```
train_channel = prefix + '/train'
validation_channel = prefix + '/validation'

sess.upload_data(path='headlines.train', bucket=bucket, key_prefix=train_channel)
sess.upload_data(path='headlines.validation', bucket=bucket, key_prefix=validation_channel)

s3_train_data = 's3://{}/{}'.format(bucket, train_channel)
s3_validation_data = 's3://{}/{}'.format(bucket, validation_channel)
```

Figure 8: Cell [4]

```
In [5]: region_name = boto3.Session().region_name
```

Figure 9: Cell [5]

```
In [6]: container = sagemaker.amazon.amazon_estimator.get_image_uri(region_name, "blazingtext", "latest")
        print('Using SageMaker BlazingText container: {} ({})'.format(container, region_name))

        Using SageMaker BlazingText container: 811284229777.dkr.ecr.us-east-1.amazonaws.com/blazingtext:latest (us-east-1)
```

Figure 10: Cell [6]

```
In [7]: s3_output_location = 's3://{}/{}/output'.format(bucket, prefix)

        bt_model = sagemaker.estimator.Estimator(container,
                                                 role,
                                                 train_instance_count=1,
                                                 train_instance_type='ml.c4.4xlarge',
                                                 train_volume_size = 30,
                                                 train_max_run = 360000,
                                                 input_mode= 'File',
                                                 output_path=s3_output_location,
                                                 sagemaker_session=sess)
```

Figure 11: Cell [7]

```
In [8]: bt_model.set_hyperparameters(mode="supervised",
                                    epochs=10,
                                    min_count=2,
                                    learning_rate=0.05,
                                    vector_dim=10,
                                    early_stopping=True,
                                    patience=4,
                                    min_epochs=5,
                                    word_ngrams=2)
```

Figure 12: Cell [8]

```
In [9]: train_data = sagemaker.session.s3_input(s3_train_data, distribution='FullyReplicated',
                                content_type='text/plain', s3_data_type='S3Prefix')
        validation_data = sagemaker.session.s3_input(s3_validation_data, distribution='FullyReplicated',
                                     content_type='text/plain', s3_data_type='S3Prefix')
        data_channels = {'train': train_data, 'validation': validation_data}
```

Figure 13: Cell [9]

```
In [10]: bt_model.fit(inputs=data_channels, logs=True)
```

Figure 14: Cell [10]

```
In [13]: text_classifier = bt_model.deploy(initial_instance_count = 1,instance_type = 'ml.m4.xlarge')
         ---------------!
```

Figure 15: Cell [11], this image was taken after several repeated runs of the code were performed which increased the cell value in the image.

```
In [14]: import nltk
nltk.download('punkt')

sentences = ["British Columbians stranded abroad feel left in the dark by government.",
             "Number of COVID-19 cases surpasses 100000 worldwide."]

# using the same nltk tokenizer that we used during data preparation for training
tokenized_sentences = [' '.join(nltk.word_tokenize(sent)) for sent in sentences]

payload = {"instances" : tokenized_sentences}

response = text_classifier.predict(json.dumps(payload))

predictions = json.loads(response)
print(json.dumps(predictions, indent=2))
```

```
[nltk_data] Downloading package punkt to /home/ec2-user/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[
  {
    "prob": [
      0.008252900093793869
    ],
    "label": [
      "__label__police"
    ]
  },
  {
    "prob": [
      0.003820229321718216
    ],
    "label": [
      "__label__no"
    ]
  }
]
```

Figure 16: Cell [12]

```
In [15]: payload = {"instances" : tokenized_sentences,
                    "configuration": {"k": 2}}

response = text_classifier.predict(json.dumps(payload))

predictions = json.loads(response)
print(json.dumps(predictions, indent=2))
```

```
[
  {
    "prob": [
      0.008252900093793869,
      0.005864045582711697
    ],
    "label": [
      "__label__police",
      "__label__new"
    ]
  },
  {
    "prob": [
      0.003820229321718216,
      0.003180541563779157
    ],
    "label": [
      "__label__no",
      "__label__new"
    ]
  }
]
```

Figure 17: Cell [13]

The screenshots of the code demonstrated above were performed on three different datasets. The first was an open source dataset entitled A Million News Headlines from the

website *Kaggle*. “This contains data of news headlines published over a period of seventeen years. Sourced from the reputable Australian news source ABC (Australian Broadcasting Corp.)” [1].

The second dataset “The AQUAINT Corpus of English News Text” was purchased from the *Linguistic Data Consortium* (LDC) database. “The AQUAINT Corpus, Linguistic Data Consortium (LDC) catalog number LDC2002T31 and ISBN 1-58563-240-6 consists of newswire text data in English, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. It was prepared by the LDC for the AQUAINT Project, and will be used in official benchmark evaluations conducted by National Institute of Standards and Technology (NIST)” [12].

The final dataset used to train and test the machine learning algorithm was a curated list created by the researcher from news articles from Canadian Broadcast Corporation website. The original design of the thesis structure was to compile a dataset of articles from early-2000s to present day headlines from the CBC news site, however due to the topical issue of novel coronavirus providing a more practical function to the algorithm model rather than a simply theoretical dataset, the topic was narrowed down to news articles regarding COVID-19.

The first two datasets are very large in volume, over a million in size for the first dataset and many thousands for the second, to the amount of training and testing data. However, the third dataset was very small in size -- twenty headlines altogether – which was in part due to the

sheer newness of the pandemic and the scope of the dataset for this pandemic being on a national scale rather than the global. The breakdown for each of the dataset sizes can be viewed in the figure below.

Iteration Model	Dataset Size
1	1,048,576
2	6,864
3	20

Table 2: Dataset Sizes

The curated dataset can be observed as follows:

- __label__We will get through this: Oilpatch hunkers down amid price plunge virus fears .
- __label__South Korea taking unprecedented steps as Italy Iran also struggle to contain COVID-19 .
- __label__COVID-19 in Quebec: Quebec up to 2840 confirmed cases premier sees encouraging signs .
- __label__British Columbians stranded abroad feel left in the dark by government .
- __label__Olympics postponement raises questions throwing athletes scheduling into disarray .
- __label__Quebec biotech firm produces a potential COVID-19 vaccine .
- __label__Coronavirus: WHO calls COVID-19 outbreak a pandemic as Italy orders most stores to close .
- __label__Quebecs first specialized COVID-19 clinic opens in Montreal .
- __label__BC Hydro says customers impacted by COVID-19 can ask for help with bill payments .

- __label__N.B. COVID-19 roundup: Province braces for next big wave of coronavirus .
- __label__Number of COVID-19 cases surpasses 100000 worldwide .
- __label__University of Alberta abandons letter grades cancels most exams amid pandemic .
- __label__Alberta radiologists bewildered and demoralized as province cancels contracts amid COVID-19 pandemic .
- __label__B.C. premier vows province will meet the challenge of COVID-19 .
- __label__We cant get home: Stuck in limbo abroad these Winnipeggers wait .
- __label__P.E.I. Premier Dennis King declares public health emergency on COVID-19 .
- __label__Canadians trapped in Morocco by COVID-19 restrictions to be evacuated this weekend: Trudeau .
- __label__Help is on the way for renters during coronavirus crisis says B.C. housing minister .
- __label__Get us out: Canadians still stranded abroad wait to hear if Ottawa will help them .
- __label__Quebec announces first death from COVID-19 confirmed cases up to 94 .

The format in which the dataset is written should be noted as being prefixed by the string "__label__" prior to the start of each headline sentence with a " ." following the end of the headline which is used as a delimiter for each line. "The BlazingText algorithm expects a single preprocessed text file with space-separated tokens. Each line in the file should contain a single sentence." [7].

Prior to the start of research, the XGBoost Algorithm on AWS was chosen to be utilized for the experiment. However, with the software development lifecycle chosen to allow room for incremental and iterative improvement to the original software plan, it was ultimately decided

that BlazingText would be a more suitable choice for the procedure. “The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms. The Word2vec algorithm is useful for many downstream natural language processing (NLP) tasks, such as sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification” [6].

BlazingText allows for both supervised and unsupervised learning. “In a supervised learning model, the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data. An unsupervised model, in contrast, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own. Semi-supervised learning takes a middle ground. It uses a small amount of labeled data bolstering a larger set of unlabeled data. And reinforcement learning trains an algorithm with a reward system, providing feedback when an artificial intelligence agent performs the best action in a particular situation.” [13]. For the use case of this research paper, it utilized the supervised multi-class text classification that the BlazingText algorithm offers. This paper also relied on the *DBPedia Text Classification BlazingText* example provided by AWS to provide guidelines and support on navigating this algorithm. BlazingText also provides unsupervised learning with its algorithm and uses Word2Vec. “Word2Vec is a neural network implementation that learns dense vector representations for words. Other deep or recurrent neural network architectures have also been proposed for learning word representations. However, they take a lot longer to train compared to Word2Vec. It directly tries to predict a word from its

neighbors, in terms of learned dense embedding vectors (considered parameters of the model), in an unsupervised way" [35].

The results of the data can be seen in the table below:

		Iteration Model		
		1	2	3
Results	*Training and Validation Accuracy*	train_accuracy: 0.122 validation_accuracy: 0.1119	train_accuracy: 0.1232 validation_accuracy: 0.1127	#train_accuracy: 0.1227 #validation_accuracy: 0.1124

	Accuracy Probability for Testing Sample 1 and Sample 2	{
		"prob": [
		0.010440291836857796,
		],
		"label": [
		"__label__police",
		]
		},
		{
		"prob": [
		0.004173622000962496,
		],
		"label": [
		"__label__no",
		]
		}
			{
			"prob": [
			0.010440291836857796
			],
			"label": [
			"__label__police"
			]
			},
			{
			"prob": [
			0.004173622000962496
			],
			"label": [
			"__label__no"
			]
			}
				{
				"prob": [
				0.008252900093793869
				],
				"label": [
				"__label__police"
				]
				},
				{
				"prob": [
				0.003820229321718216
				],
				"label": [
				"__label__no"
				]
				}

Table 3: Accuracy Results for Iteration Models

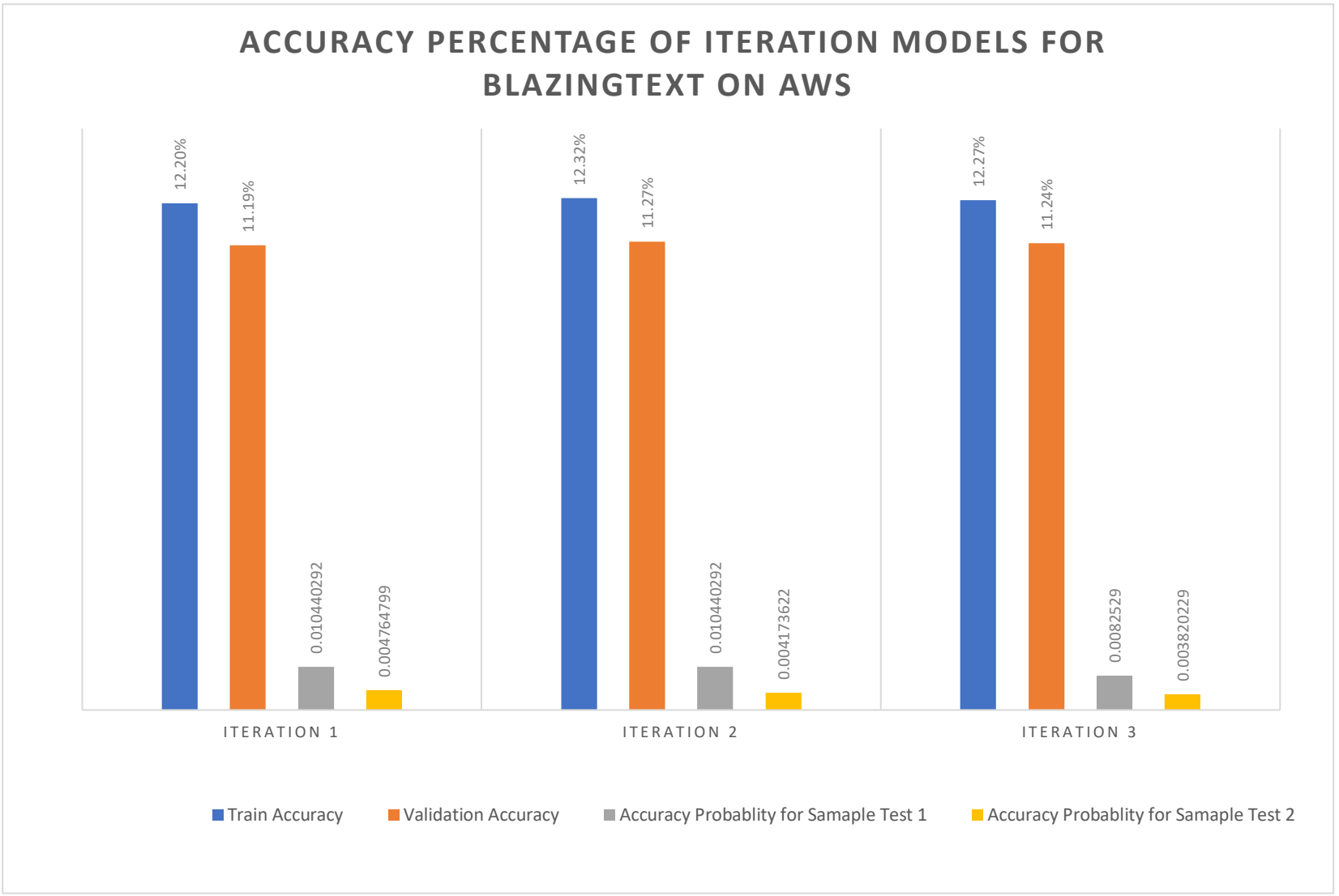


Figure 18: Chart of Results

It can be observed in the chart above that the key results of the experimentation shows that there was limited accuracy for the variables of both training accuracy and validation accuracy on all three iterations of the algorithm. There were two sample tests run that compared and tested the accuracy of the supervised learning to a predefined sentence to determine the correctness of the algorithm in detecting the sentiment. The possible causes for these results will be examined and discussed in the following chapter.

3.3 Chapter Summary

In a brief overview, the results show that there is little significance to the data produced from this experiment. There were three models used to train the BlazingText algorithm with the volume of the set sizes as follows: Model 1 having items in the dataset 1,048,576, Model 2 having items in the dataset 6,864 and Model 3 having 20 items altogether in the dataset. The models had both low training and validation accuracy which is as follows: Model 1 having training accuracy of 0.122 and validation accuracy of 0.1119, Model 2 having the training accuracy of 0.1232 and the validation accuracy of 0.1127, and Model 3 having the training accuracy of 0.1227 and the validation accuracy of 0.1124.

4. Discussion

Reflecting back upon the data that was produced during the final iteration of the iterative and incremental SDLC as documented above, there is quite a lot to unpack from these results. The following chapter aims to review the results and provide a comprehensive discussion as to what the preliminary findings may mean and their value, and the value of the software development lifecycle method chosen for the project. There will also be discussion as to how the preliminary findings relate to the literature review performed in earlier sections.

It can be observed that for the scoring of the models, Model 1 had both the lowest training accuracy of the group as well as validation accuracy. Additionally, although Model 2 and 3 increased in accuracy, the increase was not significant enough to deem the model a success. Theoretically, by increasing the total input into the training dataset, the algorithm should in turn have a higher probability of accurately detecting the sentiment of the news headline. "Since the accuracy of the word incorporation depends on a large amount of data (Big Data), it is necessary that new scalable parallel algorithms be developed to deal with this large amount of data, perhaps even billions of words. The development of scalable parallel algorithms is one of the most complex and difficult tasks…" [32]. So, with each model, the training and validation accuracy should have increased but this was not the case in this investigation.

Causes for these issues may include having noise in data that was properly cleaned during the data processing stage. Noise in data is referred to as a misleading feature which when included in document representation, will typically on average increase the classification error

[32]. Therefore, because the initial dataset had the largest quantity of headlines (the dataset sizes are as follow: Model 1 - 1,048,576 versus Model 2 – 6,864 and Model 3 – 20) to train the algorithm with, if it had any significant amount of noise this may have dramatically affected the initial training of the supervised learning. Therefore, in turn made it such that the following two dataset models were not able to cut through the noise and properly train the machine learning algorithm.

"There are several challenges that influence the prediction accuracy in text classification. Firstly, the volume of input data involved is very high hence the storage, retrieval and access to the data is a big challenge. The amount of processing units required is also high and since a single machine cannot produce such high processing capability, we need to identify a way to parallelize the algorithms and implement them in distributed computing environments. Although there are several different classifiers available there is no stable classifier available that can accurately classify all kinds of data. Before feeding the data to a classifier we need to eliminate noise based on several constraints" [32].

Other issues for the algorithm may have faced is grammatical and language nuance detections, such as that of comprehending headlines with a sarcastic tone. It is commonly known that news organisations tend to embellish, overexaggerate, or otherwise tilt their headlines to garner more readers for any particular article. The nuances of a "click-bait" headline is often presented in a manner that is understood intrinsically by the average reader but poses a challenge to computers that lack the ability to pick up on such fine subtlety. This as a result could have directly weakened training of the machine learning algorithm.

Although spelling errors are less common for news headlines, it may also have posed to be a problem that the supervised learning model was not able to overcome. “The most common and challenging problem that occurs with document classification problems is correcting the misspelt words in the corpus. Spell correction is challenging because the domain or context of the paragraph should be identified before identifying the spelling. Many spelling correction algorithms have been proposed since the early 1970’s but there are two basic principles underlying these algorithms. 1. Of all the various alternative correct spellings for a misspelled word, choose the nearest one. This demands that we have a notion of nearness or proximity between a pair of queries” [32].

Throughout the process of completing this experiment, there were many different versions and iterations performed before the official datasets and algorithms were selected for the final iteration output that is described in the results chapter. Prior to the start of performing experiments, there was research performed on the training algorithms offered by AWS which included the XGBoost and BlazingText algorithms. Due to the initial understanding of the flexibility of XGBoost, it appeared to have many benefits in utilizing this algorithm for the purposes of this research. It was also attractive because it had the ability to choose how application of the algorithm was performed - either regression or classification. However, there was a fundamental misunderstanding that XGBoost algorithms are only able to handle numerical data values. There was vigorous testing and debugging executed to ensure that the python code in SageMaker did not have any faults, but due to the basic mismatch in the input data type the program was unsuccessful in training the datasets.

After this iteration, the project moved to the BlazingText algorithm which, as the name implies, is an algorithm geared towards text and text classification. This removed the fundamental issue with the data type mismatch and was proven to be beneficial because it was geared toward the needs in the use case of this research project. “The incorporation of words made it possible to work with semantics in any application that works with a text document. Through algorithms that implement this technique, such as Word2Vec, it is possible to discover the similarity between words, paragraphs and even entire documents. However, the generation of word incorporation still has a high computational cost” [32]. The high computational requirement combined with the other issues described above may have been the holistic reason as to why the project altogether was not of statistical significance.

Turning to examining how valuable the findings of this paper are – as stated earlier in the paper, it has been defined that having a minimum accuracy of 75% would deem the results as successful. Since the results were much less than the accuracy requirement goal, it can be safely stated that the results were not statically significant. Despite the results not having value in and of itself, the process for performing the experiments provided valuable insight as to potential errors that may occur when using AWS SageMaker and recommendations on how to avoid falling into those traps.

The software development lifecycle method selected for this project has proven to be invaluable and is highly recommended for any future projects. By not having the rigid linear structure of the Waterfall SDLC which would not have been accommodating to the changing

needs or adjusting the scope of this project, there was flexibility allowed with chosen the iterative and incremental SDLC. This flexibility allowed the freedom in going back through the lifecycle to update and modify the chosen algorithms, modify the scope of the dataset for the CBC news media headlines chosen to reflect the topical pandemic of COVID-19, and these changes would not have been permissible with the Waterfall model.

The importance of data cleaning should be highlighted and emphasized in not just this experiment but all machine learning algorithms. During the various iterations of the software development lifecycle, there were different algorithms used – namely BlazingText and XGBoost – and while both algorithms are similar in terms of what they strive to achieve in terms of processing data for machine learning, they were notably different in their requirements of input data.

The way the algorithms work did not take into consideration a special set and mandated the inputs as a certain format. Getting rid of any extraneous commas and quotations marks, removing special characters and tokens, and making sure the input data file was correctly formatted, whether that be as a .csv file or a .txt file, is vitally important in ensuring that the algorithms are able to read the input data. Without being able to properly process the data will result in failure both for compiling the code throughout the research and also not providing accurate or comprehensive results. On the other hand, it should be noted that data cleaning may have some adverse effects. Notably in the XGBoost algorithm which requires csv files, wherein which commas act as delimiters, the actual data content might interfere with the delimiter requirements. Using news articles as an example, the frequent and prominent use of quotations

and numerical figures in news headlines can be observed and although this may be an attractive way to garner readers for the article, it highly disrupts the ability of an algorithm to easily and effectively input that data from a csv file into an algorithm.

By removing such tokens and characters from the csv while performing the action of cleaning the data, may result in disrupting the intent of what the headline meant to say. For example, in the headline from the CBC news website "*COVID-19 in Quebec: Quebec up to 2,840 confirmed cases, premier sees encouraging signs*" would have drastically different implications in the meaning by removing the commas within the quotation. "*COVID-19 in Quebec: Quebec up to 2840 confirmed cases premier sees encouraging signs*" strongly suggests the premier of the province views the death as encouraging and the significance of this is highly morbid, whereas the former title with the original punctuation indicates a more positive and optimistic implication despite the unfortunate deaths associated with it.

Although the known roundabout to this issue is to resolve it by wrapping the string with quotation marks which will negate the trigger of a comma to act as a delimiter in a csv file, through various stages in the testing process of this research this has been proven to be quite ineffective. "Given a document unit containing sentences or a sequence of characters, tokenization is the task of splitting up the document into meaning full words, called tokens, the major criteria to be considered while tokenizing text is understanding what special characters should be ignored during tokenization" [32]. "Tokenization generally occurs at a word level, sometimes it is difficult to define what a token exactly means. The following are some basic heuristics on which a tokenizer relies on: 1. Whitespace is generally considered a delimiter for

tokenization which is considered in most cases. 2. Punctuation and line breaks are also considered in some cases as a delimiter but after the tokenization they are removed. 3. A sequence of contiguous alphabetic characters are part of a single token; numbers may or may not be included based on the dataset" [32].

The best method to prevent failures to remove any special characters that may trigger the delimiter altogether. Therefore, it can be proposed with certainty that there must be methods created to get around this easy pitfall. Suggestions include finding ways to fix the algorithm, so it is more accurate to real world data, or perhaps creating new algorithms entirely that would avoid these traps from the source of the problem. Meanwhile, until such actions have been undertaken, it is important to ensure any extra data cleaning is performed as necessary.

5. Conclusions

"In text classification, given a set of classes and an object, the main objective is to identify to which class or set of classes the given object belongs to. Given a description d ϵ X of a document, where X is the document space; and a fixed set of "classes" C = {c1, c2, c3 … cn}. Classes are also called categories or labels. Typically, the document space is some kind of a multi-dimensional space, and the classes are most often human defined and are application specific" [32]. For this research paper, the investigation of text classification that was performed was pursued to answer the following research question: is AWS cloud-based machine learning algorithm effective for identifying sentiment of headlines for CBC news articles. The aim of this project was to query if the AWS cloud-based machine learning algorithm is efficient for identifying sentiment CBC news headlines. The effectiveness was measured by the probability in the accuracy of the machine learning algorithm. The efficiency of an algorithm was measured in a duration of seconds for how long it requires to reach a result.

The method used to reach these results were performed through various stages of testing which include the following: unit testing, integration testing, and acceptance testing. User testing was originally a testing goal for this project, however this proved ultimately to not be possible due to a limitation in the scope of the thesis timeline.

The approach in performing the scientific experiments for the research topic was performed through using Amazon Web Services via AWS SageMaker, a machine learning as a service platform. Other services that were also used on Amazon Web Services include the S3

buckets for data storage and result offloading, as well as IAM roles to ensure proper authorization to allow for full functionality of the SageMaker resources. The input data comprised of utilizing premade data sets, including those purchased from the *Linguistic Data Consortium* (LDC) database and open source datasets made available on the website *Kaggle*, as well as creating a custom dataset using headlines regarding the COVID-19 virus from the Canadian News Broadcast (CBC website). The data was cleaned by first running the datasets through several Java programs to capture the headlines from the datasets and then properly set up the layout in a manner that BlazingText on AWS would be able to accept the .txt file as a valid input format.

Although the original intent was to create a large dataset from the early-2000s to present day headlines from CBC news, the use of the coronavirus as the dataset subject had a more topical application of the machine learning algorithm. This resulted in better real-world application because at time of experimentation being carried out, COVID-19 is currently a global pandemic with a large negative effect on how it will impact how local and global societies will operate in the day-to-day and going forward into the future.

There are certain limitations that this research was unable to accomplish during the proposed time frame structure. Therefore, this has resulted in several areas not having sufficient time to be properly explored throughout this research. One example of this includes user testing which was unable to be performed due to the lengthy time required to receive the approval from the Algoma University Research Ethics Board. The typical duration of time required to submit the necessary documentations regarding the experimentation and in turn receive the proper approval from the

board exceeded the time available to complete the research project. Unfortunately, this resulted in user testing of the completed experiment not being performed. User testing would have provided a different perspective on how the experiment was run and how the data was interpreted by individuals aside from the researcher of this project. Different viewpoints allow for a more holistic approach to a subject, and it would have been interesting to see how various users interacted with the machine learning system.

Going forward, there are several recommendations for future works and how they may be improved upon from what can be gathered by the results of this paper. First and foremost, there should be a larger training dataset that must be fed into the training of the algorithm. Although one million may have been a sufficient start in this paper, the dataset size only decreased in volume which may have resulted in it hindering its ability to improve the probability in which it was able to correctly detect sentiment analysis. Additionally, more rigorous and increased forms of all the software testing would result in finding any leaks in producing efficiency of the program code. Including user testing for the software testing could also prove to be very valuable. Lastly, creating a unique supervised learning algorithm on SageMaker could pose to be an interesting challenge as it would enable further customization and specialization which could yield more accurate results.

References:

[1]"A Million News Headlines", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/therohk/million-headlines. [Accessed: 31- Mar- 2020].

[2]A. Osthoff, "Apple MacBook Pro 13 (Mid 2017, i5, without Touch Bar) Review", *NotebookCheck*, 2017. [Online]. Available: https://www.notebookcheck.net/Apple-MacBook-Pro-13-Mid-2017-i5-without-Touch-Bar-Review.234282.0.html. [Accessed: 05- Mar- 2020].

[3]J. Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison", *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128-138, 2017. Available: 10.14445/22312803/IJCTT-V48P126.

[4]"Amazon SageMaker", *AWS Developer Guide*, 2020. [Online]. Available: https://aws.amazon.com/sagemaker/. [Accessed: 05- Mar- 2020].

[5]"Amazon SageMaker BlazingText: Parallelizing Word2Vec on Multiple CPUs or GPUs | Amazon Web Services", *Amazon Web Services*, 2020. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-blazingtext-parallelizing-word2vec-on-multiple-cpus-or-gpus/. [Accessed: 01- Apr- 2020].

[6]B. Fields, "Building machine learning workflows with AWS Data Exchange and Amazon SageMaker | Amazon Web Services", *AWS Machine Learning Blog*, 2020. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/building-machine-learning-workflows-with-aws-data-exchange-and-amazon-sagemaker/. [Accessed: 31- Mar- 2020].

[7]"BlazingText Algorithm - Amazon SageMaker", *Docs.aws.amazon.com*, 2020. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html. [Accessed: 01- Apr- 2020].

[8]"Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service (S3)", *Amazon Web Services, Inc.*, 2020. [Online]. Available: https://aws.amazon.com/s3/. [Accessed: 01- Apr- 2020].

[9]"Create a Notebook Instance - Amazon SageMaker", *Docs.aws.amazon.com*, 2020. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/howitworks-create-ws.html. [Accessed: 01- Apr- 2020].

[10]D. Kong, "Science Driven Innovations Powering Mobile Product: Cloud AI vs. Device AI Solutions on Smart Device", 2017. Available: https://arxiv.org/abs/1711.07580. [Accessed 31 March 2020].

[11]"Ethics & Procedures - Algoma", *Algoma University*, 2018. [Online]. Available: https://www.algomau.ca/research/ethics-procedures/. [Accessed: 31- Mar- 2020].

[12]D. Graff, "The AQUAINT Corpus of English News Text - Linguistic Data Consortium", *Catalog.ldc.upenn.edu*, 2002. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2002T31. [Accessed: 01- Apr- 2020].

[13]I. Salian, "NVIDIA Blog: Supervised Vs. Unsupervised Learning", *NVIDIA Blog*, 2018. [Online]. Available: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/. [Accessed: 31- Mar- 2020].

[14]"IAM Roles - AWS Identity and Access Management", *Docs.aws.amazon.com*, 2020. [Online]. Available: https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles.html. [Accessed: 01- Apr- 2020].

[15]J. Cadle, *Developing Information Systems - Practical guidance for IT professionals*. Swindon: BCS Learning & Development Ltd, 2014, p. 115.

[16]I. Foster and D. Gannon, *Cloud Computing for Science and Engineering*. Cambridge: The MIT Press, 2017, pp. 21-28, 191-217.

[17]J. Furbush, "Machine learning: A quick and simple definition", *O'Reilly Media*, 2018. [Online]. Available: https://www.oreilly.com/content/machine-learning-a-quick-and-simple-definition/. [Accessed: 31- Mar- 2020].

[18]J. Hurwitz and D. Kirsch, "What is machine learning?", *Ibm.com*, 2018. [Online]. Available: https://www.ibm.com/topics/machine-learning. [Accessed: 31- Mar- 2020].

[19]J. Jackovich and R. Richards, *Machine learning with AWS*. Birmingham, UK: Packt Publishing, 2018.

[20]J. Mueller and L. Massaron, *Machine Learning for dummies*. Hoboken: John Wiley & Sons, 2016.

[21]K. Choi, S. Aich and H. Kim, "A Machine Learning Approach to Predict Happiness Based on Sentiment Analysis of Twitter Data", 2018. Available: http://www.dbpia.co.kr/Article/NODE07485361. [Accessed 31 March 2020].

[22]M. Garcia-Ruiz, "Updating your Methods Section", Algoma University, 2020.

[23]M. Harrison, *Machine Learning Pocket Reference: Working With Structured Data in Python*. Sebastopol: O'Riley, 2019.

[24]M. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects", *Science*, vol. 349, no. 6245, pp. 255-260, 2015. Available: 10.1126/science.aaa8415.

[25]M. Xiu, Z. Jiang and B. Adams, "An Exploratory Study on Machine-Learning Model Stores", 2020. Available: https://arxiv.org/abs/1905.10677v2.

[26]"Machine Learning - Evaluate - ML Studio (classic) - Azure", *Docs.microsoft.com*, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-evaluate. [Accessed: 31- Mar- 2020].

[27]"Machine Learning - Score", *Docs.microsoft.com*, 2020. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-score. [Accessed: 31- Mar- 2020].

[28]Maersk Line - The Innovation & Transformation of Maersk Line. Leading Practices from the Outperformers case study research paper - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/traditional-Waterfall-versus-Agile-Model-ix_fig6_287816444 [accessed 2 Mar, 2020]

[29]N. Castle, "Regression vs. Classification Algorithms", *Blogs.oracle.com*, 2018. [Online]. Available: https://blogs.oracle.com/datascience/regression-vs-classification-algorithms. [Accessed: 01- Apr- 2020].

[30]N. Gift, *Pragmatic AI*. Addison-Wesley Professional, 2018.

[31]O. DeMasi, K. Kording and B. Recht, "Meaningless comparisons lead to false optimism in medical machine learning", 2017. Available: https://arxiv.org/abs/1707.06289v1.

[32]R. Nandan Immaneni, "An Efficient Approach to Machine Learning Based Text Classification Through Distributed Computing", Masters, California State University, Long Beach, 2015.

[33]S. Bringsjord and N. Sundar Govindarajulu, "Artificial Intelligence (Stanford Encyclopedia of Philosophy)", *Plato.stanford.edu*, 2018. [Online]. Available: https://plato.stanford.edu/entries/artificial-intelligence/. [Accessed: 31- Mar- 2020].

[34]S. Gupta and V. Khare, "Enhanced text classification and word vectors using Amazon SageMaker BlazingText | Amazon Web Services", *AWS Machine Learning Blog*, 2018. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/enhanced-text-classification-and-word-vectors-using-amazon-sagemaker-blazingtext/. [Accessed: 31- Mar- 2020].

[35]S. Gupta and V. Khare, "Amazon SageMaker BlazingText: Parallelizing Word2Vec on Multiple CPUs or GPUs", *AWS Machine Learning Blog*, 2018. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/enhanced-text-classification-and-word-vectors-https://aws.amazon.com/blogs/machine-learning/amazon-sagemaker-blazingtext-parallelizing-word2vec-on-multiple-cpus-or-gpus/. [Accessed: 31- Mar- 2020].

[36]S. Russell and P. Norvig, *Artificial intelligence*, 3rd ed. Upper Saddle River: Pearson, 2010, pp. 5-28, 64- 157.

[37]T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing", 2018. Available: https://arxiv.org/pdf/1708.02709.pdf.

[38]"Train - ML Studio (classic) - Azure", *Docs.microsoft.com*, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-train. [Accessed: 31- Mar- 2020].

[39]"Train a Model with Amazon SageMaker - Amazon SageMaker", AWS Documentation, 2019. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html. [Accessed: 01- Dec- 2019].

[40]W. Caesarendra et al., "An AWS Machine Learning-Based Indirect Monitoring Method for Deburring in Aerospace Industries Towards Industry 4.0", *Applied Sciences*, vol. 8, no. 11, p. 2165, 2018. Available: 10.3390/app8112165.

[41]W. Van Casteren, "The Waterfall Model and the Agile Methodologies : A comparison by project characteristics", 2020. Available: https://www.researchgate.net/publication/313768756_The_Waterfall_Model_and_the_Agile_Methodologies_A_comparison_by_project_characteristics. [Accessed 1 March 2020].

[42]"What is Natural Language Processing?", *sas.com*, 2020. [Online]. Available: https://www.sas.com/en_ca/insights/analytics/what-is-natural-language-processing-nlp.html. [Accessed: 31- Mar- 2020].

[43]"XGBoost Algorithm", *AWS Developer Guide*, 2020. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html. [Accessed: 05- Mar- 2020].

Appendix

Step 1) Data Cleaning – Collecting headlines from dataset

```
package thesis;
import java.io.BufferedWriter;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.Scanner;

public class HeadlinesFile {
	public static void main(String args[]) throws IOException {
	    PrintWriter out = null;
	    FileReader in = null;
	    try {
	      out = new PrintWriter(new BufferedWriter(new FileWriter("output.txt", true)));
	      in = new FileReader("output.txt");


	      Scanner scanner = new Scanner(in);

	      while (scanner.hasNextLine()) {
	    		String line = scanner.nextLine();
	        if(line.contains("<HEADLINE>")) {
	    		out.write(scanner.nextLine() + " \n");
	        }


	      }
	    }finally {
	      if (in != null) {
	        in.close();
	      }
	      if (out != null) {
	        out.close();
	      }
	    }
	  }
}
```

Step 2) Data Cleaning – Removing unwanted characters and strings from dataset

```
package thesis;
import java.io.BufferedWriter;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.Scanner;

public class RemoveHT {
	public static void main(String args[]) throws IOException {
		 FileReader in = null;
	    FileWriter out = null;


	    try {
	      in = new FileReader("input.txt");
	      out = new FileWriter("output.txt");

	      Scanner scanner = new Scanner(in);

	      while (scanner.hasNextLine()) {
			String line = scanner.nextLine();
	        if(!line.contains("&HT;")) {
			out.write(line + " \n");
	        }


	      }
	    }finally {
	      if (in != null) {
	        in.close();
	      }
	      if (out != null) {
	        out.close();
	      }
	    }
	  }
}
```

Step 3) Data Cleaning – Formatting the .txt files to include "__label__" and " ." per BlazingText requirements

```
package thesis;
import java.io.*;
import java.util.Scanner;

public class CopyFile {

  public static void main(String args[]) throws IOException {
    FileReader in = null;
    FileWriter out = null;


    try {
      in = new FileReader("input.txt");
      out = new FileWriter("output.txt");

      Scanner scanner = new Scanner(in);

      while (scanner.hasNextLine()) {
        String line = scanner.nextLine();
        if(line.charAt(0) != ' ')
                out.write("__label__" + line + " .\n");
      }

    }finally {
      if (in != null) {
        in.close();
      }
      if (out != null) {
        out.close();
      }
    }
  }
}
```

Step 4) Machine Learning – AWS SageMaker Python Code

```
import sagemaker
import boto3
import pandas as pd
from sagemaker import get_execution_role
import json

sess = sagemaker.Session()

role = get_execution_role()
print(role)

region = boto3.Session().region_name
bucket = 'dluitel'
print(bucket)
prefix = 'sagemaker/blazingtext'
bucket_path = 'https://s3-{}.amazonaws.com/{}'.format(region,bucket)
print(bucket_path)


import io
import boto3
import random

def data_split(FILE_DATA, FILE_TRAIN, FILE_VALIDATION, FILE_TEST,
PERCENT_TRAIN, PERCENT_VALIDATION, PERCENT_TEST):
    data = [l for l in open(FILE_DATA, 'r')]
    train_file = open(FILE_TRAIN, 'w')
    valid_file = open(FILE_VALIDATION, 'w')
    tests_file = open(FILE_TEST, 'w')

    num_of_data = len(data)
    num_train = int((PERCENT_TRAIN/100.0)*num_of_data)
    num_valid = int((PERCENT_VALIDATION/100.0)*num_of_data)
    num_tests = int((PERCENT_TEST/100.0)*num_of_data)

    data_fractions = [num_train, num_valid, num_tests]
    split_data = [[],[],[]]

    rand_data_ind = 0
```

```
    for split_ind, fraction in enumerate(data_fractions):
        for i in range(fraction):
            rand_data_ind = random.randint(0, len(data)-1)
            split_data[split_ind].append(data[rand_data_ind])
            data.pop(rand_data_ind)

    for l in split_data[0]:
        train_file.write(l)

    for l in split_data[1]:
        valid_file.write(l)

    for l in split_data[2]:
        tests_file.write(l)

    train_file.close()
    valid_file.close()
    tests_file.close()

def write_to_s3(fobj, bucket, key):
    return
boto3.Session(region_name=region).resource('s3').Bucket(bucket).Object(key).upload_fileobj(fo
bj)

def upload_to_s3(bucket, channel, filename):
    fobj=open(filename, 'rb')
    key = prefix+'/'+channel
    url = 's3://{}/{}/{}'.format(bucket, key, filename)
    print('Writing to {}'.format(url))
    write_to_s3(fobj, bucket, key)


import urllib.request

FILE_DATA = 'uci-headlines'

#split the downloaded data into train/test/validation files
FILE_TRAIN = 'headlines.train'
FILE_VALIDATION = 'headlines.validation'
```

```
FILE_TEST = 'headlines.test'

PERCENT_TRAIN = 70
PERCENT_VALIDATION = 15
PERCENT_TEST = 15

data_split(FILE_DATA, FILE_TRAIN, FILE_VALIDATION, FILE_TEST,
PERCENT_TRAIN, PERCENT_VALIDATION, PERCENT_TEST)


train_channel = prefix + '/train'
validation_channel = prefix + '/validation'

sess.upload_data(path='headlines.train', bucket=bucket, key_prefix=train_channel)
sess.upload_data(path='headlines.validation', bucket=bucket, key_prefix=validation_channel)

s3_train_data = 's3://{}/{}'.format(bucket, train_channel)
s3_validation_data = 's3://{}/{}'.format(bucket, validation_channel)


region_name = boto3.Session().region_name

container = sagemaker.amazon.amazon_estimator.get_image_uri(region_name, "blazingtext",
"latest")
print('Using SageMaker BlazingText container: {} ({})'.format(container, region_name))

s3_output_location = 's3://{}/{}/output'.format(bucket, prefix)

bt_model = sagemaker.estimator.Estimator(container,
                                         role,
                                         train_instance_count=1,
                                         train_instance_type='ml.c4.4xlarge',
                                         train_volume_size = 30,
                                         train_max_run = 360000,
                                         input_mode= 'File',
                                         output_path=s3_output_location,
                                         sagemaker_session=sess)


bt_model.set_hyperparameters(mode="supervised",
```

```
                    epochs=10,
                    min_count=2,
                    learning_rate=0.05,
                    vector_dim=10,
                    early_stopping=True,
                    patience=4,
                    min_epochs=5,
                    word_ngrams=2)


train_data = sagemaker.session.s3_input(s3_train_data, distribution='FullyReplicated',
                    content_type='text/plain', s3_data_type='S3Prefix')
validation_data = sagemaker.session.s3_input(s3_validation_data, distribution='FullyReplicated',
                     content_type='text/plain', s3_data_type='S3Prefix')
data_channels = {'train': train_data, 'validation': validation_data}

bt_model.fit(inputs=data_channels, logs=True)
text_classifier = bt_model.deploy(initial_instance_count = 1,instance_type = 'ml.m4.xlarge')

import nltk
nltk.download('punkt')
sentences = ["British Columbians stranded abroad feel left in the dark by government.",
          "Number of COVID-19 cases surpasses 100000 worldwide."]
# using the same nltk tokenizer that we used during data preparation for training
tokenized_sentences = [' '.join(nltk.word_tokenize(sent)) for sent in sentences]

payload = {"instances" : tokenized_sentences}

response = text_classifier.predict(json.dumps(payload))

predictions = json.loads(response)
print(json.dumps(predictions, indent=2))

payload = {"instances" : tokenized_sentences,
        "configuration": {"k": 2}}

response = text_classifier.predict(json.dumps(payload))

predictions = json.loads(response)
print(json.dumps(predictions, indent=2))
```

Article 1: https://www.cbc.ca/news/canada/calgary/calgary-oilpatch-hunkers-down-1.5495162

Calgary

'We will get through this': Oilpatch hunkers down amid price plunge, virus fears

Companies scrutinize priorities as they prepare for 'extraordinary' times

Tony Seskus · CBC News · Posted: Mar 12, 2020 3:29 PM MT | Last Updated: March 12

A pumpjack pulls oil out of the ground in central Alberta. The benchmark price for North American crude closed Thursday at $31.50 US per barrel, down $1.48. (Kyle Bakx/CBC)

On Monday morning, as oil prices melted down around the world, Grant Fagerheim met with employees at Whitecap Resources to discuss their health and safety amid fears of the spread of the novel coronavirus.

"We pulled the entire staff together and had these conversations," Fagerheim, the chief executive of the mid-sized producer, told the *Calgary Eyeopener* on Thursday.

"We're hopeful that we can calm the waters here. Again, as I say, we will get through this. We just have to make decisions in a very organized manner versus being too reactive."

Like its peers around the sector, Whitecap is being challenged on two fronts: the crude price war that's broken out between Russia and Saudi Arabia, and the spread of the virus, which is rattling markets to the core.

But companies must also manage the practical health concerns of staff — a situation highlighted by news a daycare in Suncor Energy's downtown headquarters has been closed after a child tested positive for COVID-19, the respiratory illness caused by the

Article 2: https://www.cbc.ca/news/world/coronavirus-clusters-outside-china-1.5473011

World

South Korea taking 'unprecedented' steps as Italy, Iran also struggle to contain COVID-19

President Moon Jae-in puts South Korea on highest alert as number of infections soar

The Associated Press · Posted: Feb 23, 2020 5:54 AM ET | Last Updated: February 23

A member of the medical team takes a rest outside a hospital in Daegu, South Korea, on Sunday. (Im Hwa-young/Yonhap via The Associated Press)

Some of the latest numbers:

- **Wuhan doctor who postponed wedding to fight outbreak dies from virus.**
- **Iran has 43 cases of COVID-19, eight deaths, and 15 new cases on Sunday.**
- **Italy has 132 cases and three deaths.**
- **South Korea has 602 cases and six deaths.**
- **China reports a total of 76,936 cases, including 648 new cases on Saturday, up from 387 a day earlier.**
- **Britain now has 13 cases, with four former cruise ship passengers testing positive.**

Article 3: https://www.cbc.ca/news/canada/montreal/covid-19-in-quebec-1.5514000

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

Montreal

COVID-19 in Quebec: Province up to 2,840 confirmed cases, but premier sees encouraging signs

Public health director says it's still too early to ease restrictions

Colin Harris, Marilla Steuter-Martin · CBC News ·
Posted: Mar 29, 2020 7:48 AM ET | Last Updated: March 29

'Public health authorities are telling us that our efforts are paying off, so don't give up,' Premier François Legault said at his daily news conference in Quebec City. (Jacques Boissinot/The Canadian Press)

22 comments

- ***Quebec has 2,840 confirmed cases and 22 deaths attributable to COVID-19. Seventy-two are in intensive care.***
- ***Montreal, which is under a local state of emergency, has about half of those cases with 1,361.***
- ***A drive-thru testing opened in Côte Saint-Luc today. Pedestrians are asked to use the clinic at Place-des-Festivals instead.***
- ***Travel into certain regions with vulnerable populations is being restricted by police checkpoints.***
- ***A first case has been reported in Nunavik, Que., according to the regional health authority.***
- ***The Quebec government has made physical distancing instructions available in more than a dozen languages, including Arabic, Creole and Yiddish.***

Article 4: https://www.cbc.ca/news/canada/british-columbia/british-columbians-stranded-abroad-feel-left-in-the-dark-by-government-1.5502379

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

British Columbia

British Columbians stranded abroad feel left in the dark by government

Canadians in countries with closed borders might not be able to return for awhile, says Global Affairs

Joel Ballard · CBC News · Posted: Mar 18, 2020 7:04 PM PT | Last Updated: March 18

Many British Columbians stuck abroad say it's been difficult to connect with their local embassies. (Alex Scobie, Dan Harney, Oliver Chapman)

comments

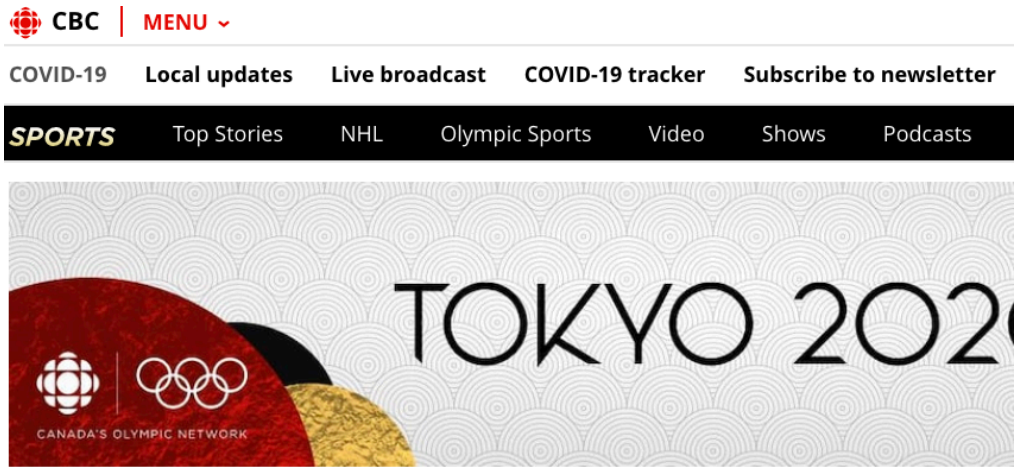
In countries across the world, British Columbians travelling abroad are trying to follow the advice of the prime minister to come home.

But many are finding it difficult to follow Trudeau's direction, as flight cancellations and closed borders foil their attempts to get back to Canada.

- **Trudeau says Canadians will be stranded abroad. Here's what travellers need to know**

CBC News spoke with many British Columbians who are stuck abroad about the different roadblocks they're facing and the lack of support they say they're receiving.

Article 5: https://www.cbc.ca/sports/olympics/postponement-raises-more-questions-throwing-olympians-rigid-scheduling-into-disarray-1.5507990



Olympics

Olympics postponement raises questions, throwing athletes' scheduling into disarray

'What is the start date? What does qualification look like,' asks wrestler Erica Wiebe

Devin Heroux · CBC Sports ·
Posted: Mar 24, 2020 12:00 PM ET | Last Updated: March 24

Empty seats provide the background as Canadian wrestler Erica Wiebe steps on to the wrestling mat for competition at an Olympic qualifying tournament in Ottawa earlier this month. (Adrian Wyld/Canadian Press)

105 comments

It's still going to be called Tokyo 2020, but the Olympics and Paralympics are now, most likely, going to be happening in the summer of 2021.

Article 6: https://www.cbc.ca/news/canada/montreal/biotech-quebec-city-coronavirus-1.5497043

Montreal

Quebec biotech firm produces a potential COVID-19 vaccine

'Extremely fast' turnaround, but months of tests required to verify effectiveness in humans

John MacFarlane · CBC News ·
Posted: Mar 13, 2020 3:24 PM ET | Last Updated: March 13

Quebec City-based Medicago uses plants to mass-produce vaccines. (Medicago)

A biotechnology company in Quebec City that previously developed and mass produced a flu vaccine says it has produced a potential vaccine for COVID-19.

This is not yet a coronavirus vaccine — at this stage it is a "vaccine candidate" that could still prove to be ineffective during several steps and test phases.

But as scientists around the world race to find a vaccine, the 20-day turnaround by the company, Medicago, is notable: vaccine candidates are typically developed on a scale of months or more.

"We can produce a vaccine candidate very quickly once a pandemic is declared," said Nathalie Landry, Medicago's executive vice-president of scientific and medical affairs.

"We can scale up very quickly and produce a large number of doses. If you think of influenza, for example, it takes anything from four to six months before vaccine doses are produced."

Article 7: https://www.cbc.ca/news/world/coronavirus-pandemic-1.5493411

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

World

Coronavirus: WHO calls COVID-19 outbreak a pandemic as Italy orders most stores to close

Italy on front line, WHO cautions other countries 'will be in that situation soon'

The Associated Press · Posted: Mar 11, 2020 7:03 AM ET | Last Updated: March 11

Despite saying the coronavirus outbreak is now a pandemic, the World Health Organization says that does not change what countries should do. 2:47

1464 comments

The latest:

- ***WHO describes outbreak as a pandemic as case numbers top 118,000 in 114 countries.***
- ***U.S. is suspending all travel from Europe for the next 30 days, but not the U.K.***
- ***Italy reports more than 12,000 cases of COVID-19, with more than 800 deaths.***
- ***Trudeau announces $1-billion fund for COVID-19 response.***
- ***Guatemala bars Europeans, starting Thursday.***
- ***G7 meetings in Pittsburgh will be held by video conference.***
- ***Read more about how Canadians are being urged to help 'flatten the curve' of COVID-19.***
- ***Watch tonight, on a special edition of The National: What you need to know about the coronavirus pandemic. How people around the world and here at home are bracing, and our panel of doctors answers your questions.***

Article 8: https://www.cbc.ca/news/canada/montreal/coronavirus-covid-19-quebec-1.5490706

Montreal

Quebec's first specialized COVID-19 clinic opens in Montreal

Province identified fifth presumptive case of coronavirus Monday

CBC News · Posted: Mar 09, 2020 7:49 AM ET | Last Updated: March 9

A security guard wears a protective mask in front of the new COVID-19 clinic which is now open at the site of the former Hôtel-Dieu Hospital. (Ryan Remiorz/The Canadian Press)

2 comments

The first of three planned specialized COVID-19 screening clinics has opened in the shuttered emergency room of Montreal's Hôtel-Dieu Hospital Monday — the same day that Quebec's fifth presumptive case of an infection has been identified.

Quebec's Health Ministry says this fifth presumptive case was a man who had been recently travelling. A sample has been sent to Canada's National Microbiology Laboratory in Winnipeg for confirmation.

Health Minister Danielle McCann said rather than just showing up at one of the new screening clinics, people who feel they have symptoms of COVID-19 should call Info-Santé 811 first to speak with a nurse.

At that point, if the nurse feels they may have the virus, they will book an appointment for the caller to be screened at one of the specialized clinics.

Article 9: https://www.cbc.ca/news/canada/british-columbia/bc-hydro-coronavirus-defer-payments-1.5494246

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

British Columbia

BC Hydro says customers impacted by COVID-19 can ask for help with bill payments

Corporation says some people could be eligible for a crisis grant of up to $600

CBC News · Posted: Mar 11, 2020 2:27 PM PT | Last Updated: March 11

BC Hydro says it has been monitoring COVID-19 since January and is encouraging customers impacted by the virus to contact the corporation for help with bill payments. (Maggie MacPherson/CBC)

0 comments

BC Hydro customers who have suffered financial hardship due to the novel coronavirus may qualify for payment relief.

In a statement to CBC, the Crown corporation says it will work with customers who are having trouble paying their bill to either defer payments or arrange a flexible payment plan.

Customers may also be eligible to access BC Hydro's Customer Crisis Fund, which provides access to grants of up to $600 to pay their bill.

To qualify for one of these grants, a customer must be facing temporary financial hardship due to job loss, injury, illness or the death of a family member.

As of March 11, a total of 46 positive cases of coronavirus have been identified in the province.

Article 10: https://www.cbc.ca/news/canada/new-brunswick/covid-19-coronavirus-outbreak-premier-blaine-higgs-1.5504334

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

New Brunswick

N.B. COVID-19 roundup: Province braces for 'next big wave' of coronavirus

Premier Blaine Higgs and Dr. Jennifer Russell say returning travellers who have no symptoms pose risk

Bobbi-Jean MacKinnon, Elizabeth Fraser · CBC News ·
Posted: Mar 20, 2020 2:08 PM AT | Last Updated: March 21

The province is urging returning travellers to have someone drop off a vehicle at the airport for them, rather than being picked up by someone. (Photo: Mike Heenan/CBC)

77 comments

Extra resources are being added to the Tele–Care 811 line to reduce wait times as New Brunswick faces the "next big wave" of COVID-19 potentially entering the province, government officials announced Friday.

Premier Blaine Higgs was referring to residents returning from travel outside Canada, stressing the importance of them following the self-isolation rules as soon as they get off their planes.

New Brunswick declared a state of emergency Thursday in response to the outbreak, giving the government broad powers to enforce business closures and social distancing to prevent the spread of the virus.

Article 11: https://www.cbc.ca/news/world/covid-19-coronavirus-march-6-1.5488009

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

World

Number of COVID-19 cases surpasses 100,000 worldwide

Bulk of new cases in epidemic continues to shift from China to other countries

Thomson Reuters · Posted: Mar 06, 2020 6:57 AM ET | Last Updated: March 6

From hygiene tips to conversations you might not have considered having, here's what you can do to get ready for an outbreak, according to Dr. Maria Van Kerkhove of the WHO Health Emergencies Programme. 1:57

1023 comments

The latest:

- ***COVID-19 cases surpass 100,000 globally.***
- ***21 people, mostly crew members, test positive for coronavirus on Grand Princess cruise ship.***
- ***Canada's chief public health officer urges travellers to "think twice" before going on a cruise ship.***
- ***WHO chief concerned by uptick in cases in low-income countries with weaker health systems.***
- ***Italy says death toll up to 197, with 4,636 cases. Iran also reports uptick, with 124 deaths and 4,747 confirmed cases.***
- ***Total number of cases in Canada rises to 54.***
- ***Read more about how Canada will cope with community transmission of the coronavirus.***

Article 12: https://www.cbc.ca/news/canada/edmonton/university-of-alberta-covid-exams-grades-1.5504807

Edmonton

University of Alberta abandons letter grades, cancels most exams amid pandemic

'I did work quite hard and now it's all being thrown away'

Wallis Snowdon · CBC News ·
Posted: Mar 20, 2020 12:53 PM MT | Last Updated: March 20

Students walking on the University of Alberta campus. (Tricia Kindleman/CBC)

The University of Alberta has temporarily abandoned a letter grade system and has "strongly encouraged" professors to cancel final exams this spring in light of heightened concerns around the spread of COVID-19.

The measures were unanimously approved Thursday during a special meeting of the executive of the U of A's general faculties council.

Undergraduate and graduate students enrolled in the winter semester will be assigned one of three marks on their transcripts: credit, no credit or incomplete. The grades will carry no weight in calculating a student's grade-point average.

Exemptions to the grading scheme may be established by the deans. The deadline for students to withdraw from classes will also be extended.

Article 13: https://www.cbc.ca/news/canada/calgary/alberta-radiology-contracts-termination-notice-pandemic-1.5499535

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

Calgary

Alberta radiologists 'bewildered and demoralized' as province cancels contracts amid COVID-19 pandemic

'It just feels like more bad-faith bargaining from government. We're really not sure why they've done this.'

Robson Fletcher · CBC News ·
Posted: Mar 16, 2020 4:31 PM MT | Last Updated: March 16

A thoracic radiologist points to an image showing a case of lung cancer in this file photo. (Carolyn Ray/CBC)

484 comments

Radiologists are "bewildered and demoralized" after the Alberta government gave notice on Friday, amid the global COVID-19 pandemic, that it would be terminating contracts signed months earlier with three of the largest diagnostic-imaging providers in the province, says the head of their professional association.

Dr. Robert Davies, president of the Alberta Society of Radiologists, which represents more than 90 per cent of radiologists in the province, said the notice of termination came as a shock to members.

Article 14: https://www.cbc.ca/news/canada/british-columbia/premier-b-c-reaction-march-13-1.5497589

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

British Columbia

B.C. premier vows province will meet the challenge of COVID-19

John Horgan says the province and federal government are united in meeting the virus' challenge

Roshini Nair · CBC News · Posted: Mar 13, 2020 4:32 PM PT | Last Updated: March 13

Premier John Horgan addressed reporters with an update on the province's response to the coronavirus. (Mike McArthur/CBC)

43 comments

British Columbia's premier, John Horgan, called the spread of COVID-19 across the province and country "uncharted territory" in an update following a meeting of first ministers on Friday.

Horgan noted the meeting, which had to be held via teleconference due to Prime Minister Justin Trudeau being under self-isolation after his wife Sophie tested positive for coronavirus, had a "unanimity of purpose."

"This is an unprecedented period in our history," Horgan said. "But it is a challenge we will take up together."

Article 15: https://www.cbc.ca/news/canada/manitoba/winnipeggers-covid-stuck-abroad-1.5500769

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

Manitoba

'We can't get home': Stuck in limbo abroad, these Winnipeggers wait

Uncertainty and confusion prevail as COVID-19 limits international travel

Austin Grabish · CBC News ·
Posted: Mar 17, 2020 4:30 PM CT | Last Updated: March 17

Winnipeg artist Franklin Fernando isn't sure when he'll be able to return home from Sri Lanka, where coronavirus-related restrictions are preventing travel. (Franklin Fernando)

On the streets of Lima, Peru, there are armed soldiers warning people to keep their distance from each other.

The country's borders have closed for at least 15 days and Winnipeggers on a gap year trip in the country are now stranded.

"They walked up to me and they told us to separate from each other, just to space out," said Piper Larsen, 23 in a FaceTime interview explaining the military presence Tuesday.

Larsen has been travelling through South America since early January.

She has been in Peru since March 6, with two other friends who are celebrating graduating from university. When they arrived in that country, there was only one reported case of the coronavirus, she said.

Article 16: https://www.cbc.ca/news/canada/prince-edward-island/pei-chief-public-health-emergency-update-covid-19-1.5499470

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

PEI

P.E.I. Premier Dennis King declares 'public health emergency' on COVID-19

King announces $25M COVID-19 emergency contingency fund for Islanders financially affected

Sam Juric · CBC News ·
Posted: Mar 16, 2020 4:49 PM AT | Last Updated: March 16

Morrison and King were accompanied by provincial cabinet ministers for the announcement. (Julien Lecacheur/CBC-Radio Canada)

3 comments

In Monday's second press briefing of the day, P.E.I. Premier Dennis King, via conference call, declared a public health emergency and is directing all provincial government employees who can work from home to do so for the next two weeks.

King was accompanied by P.E.I.'s Chief Public Health Officer Dr. Heather Morrison and some of the province's cabinet ministers.

"A public health emergency is declared when either a health emergency exists or is imminent. And both are true in this case. Now is the time to do so and it allows us to make the decisions that we may need to in the days ahead," Morrison said.

Article 17: https://www.cbc.ca/news/politics/covid-19-coronavirus-pandemic-morocco-1.5504550

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

Politics

Canadians trapped in Morocco by COVID-19 restrictions to be evacuated this weekend: Trudeau

Tens of thousands are stranded abroad as nations move to slow spread of COVID-19

John Paul Tasker · CBC News ·
Posted: Mar 20, 2020 1:05 PM ET | Last Updated: March 20

Passengers line up to board one of the few flights out of Morocco in Marrakech on Thursday. (Jessica Blough/The Associated Press)

189 comments

Prime Minister Justin Trudeau said today that a flight has been arranged to bring home Canadians stranded in Morocco, as the COVID-19 pandemic continues to wreak havoc on the transport sector.

"We're in discussion with Canadian airlines to help Canadians stranded abroad come home," Trudeau said this morning from outside Rideau Cottage in Ottawa, where he remains in self-isolation after his wife tested positive for COVID-19. "We will have more details to share but the first flight will be picking up Canadians from Morocco this weekend."

Article 18: https://www.cbc.ca/news/canada/british-columbia/bc-housing-coronavirus-1.5505626

British Columbia

'Help is on the way' for renters during coronavirus crisis, says B.C. housing minister

Officials also lay out plan for vulnerable people on Vancouver's DTES, from meal delivery to possible moves

Lisa Johnson · CBC News · Posted: Mar 21, 2020 9:32 AM PT | Last Updated: March 21

B.C. Housing Minister Selina Robinson said Saturday the government is looking at broader measures to prevent evictions, but did not provide specifics. (CBC)

67 comments

With less than two weeks until rent is due, and thousands of British Columbians losing their jobs due to the COVID-19 crisis, many are wondering how they'll keep their homes.

"I know that many people are worried about how they are going to make ends meet, put food on the table, and pay rent" by April 1, said B.C. Housing Minister Selena Robinson at a press conference Saturday morning.

"Help is on the way" for renters, she promised.

However, Robinson offered no details today on what form that relief might take, saying Finance Minister Carole James would make an announcement early next week.

Article 19: https://www.cbc.ca/news/business/canadians-travellers-stranded-abroad-covid-19-closed-borders-ecuador-honduras-1.5505121

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

Business

'Get us out': Canadians still stranded abroad wait to hear if Ottawa will help them

Those stranded abroad face closed borders and no sign of a way out

Sophia Harris · CBC News ·
Posted: Mar 21, 2020 4:00 AM ET | Last Updated: March 21

Paul Latendresse and wife, Diane Villeneuve, showed up at the Casablanca airport in Morocco Saturday in hopes of getting a flight back to Canada, but were unsuccessful. (submitted by Paul Latendresse)

795 comments

"Bring us home." That's Shirley Mancino's message to Justin Trudeau after the prime minister pleaded earlier this week for Canadians abroad to return to Canada during the COVID-19 pandemic.

Mancino, 74, and husband, Michael Clement, 75, left their home in Westport, Ont., in January to spend the winter in Cuenca, Ecuador. They were set to fly home on April 8, but their plan was dashed when Ecuador closed its borders on March 16 to help stop the spread of COVID-19.

"The president of Ecuador just slammed the door completely," said Clement from the couple's rented apartment in Cuenca. "Meanwhile, Trudeau is telling us, 'Why don't you come home?'"

Article 20: https://www.cbc.ca/news/canada/montreal/covid-19-march18-1.5501239

CBC | MENU

COVID-19 Local updates Live broadcast COVID-19 tracker Subscribe to newsletter

news Top Stories Local The National Opinion World Canada

Montreal

Quebec announces first death from COVID-19, confirmed cases up to 94

But authorities plead with Quebecers not to despair — don't give up, Arruda says

Jonathan Montpetit, Laura Marchand · CBC News ·
Posted: Mar 18, 2020 6:32 AM ET | Last Updated: March 18

'The health-care network is ready,' Premier François Legault said Wednesday. (Jacques Boissinot/The Canadian Press)

7 comments

- Premier François Legault announced the **first death** in Quebec from COVID-19: an elderly person in the Lanaudière region.
- There are **94 confirmed cases** in Quebec, up from 50 on Monday. **Four** people are in intensive care.
- Call **1-877-644-4545** if you think you have COVID-19 symptoms instead of 811.
- Quebec's director of public health, Dr. Horacio Arruda, is **asking Quebecers not to wear masks**. He said it does not prevent the spread of COVID-19, and masks should be reserved for health care workers.
- Someone with a confirmed case of COVID-19 took a **shuttle from the airport** to an airport parking lot on **March 8.**
- Canada and the United States will **close the border to non-essential travel** while allowing some commercial traffic to continue.
- Air Transat and Porter are **suspending flights**.