A Project Report on

Early Warning System using Global News For Investors and Creditors

Submitted in partial fulfilment for award of degree of

MBA

In **Business Analytics**

Submitted by

SNEHA P TIWARI

R19MBA08

Under the Guidance of

AKSHAY KULKARNI

Lead Data Scientist

40 under 40 Data Scientist | Speaker | Author | AI Guest

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

October 2020

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

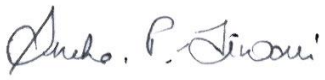
I, Sneha P Tiwari hereby declare that I have completed the project work towards the MBA in Business Analytics at, REVA University on the topic entitled **Early Warning System for Credit Risk Management** under the supervision of Akshay Kulkarni, Lead Data Scientist, This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year 2020.

Place: Bengaluru

Date:

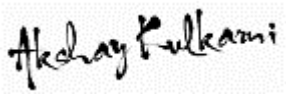
Name of the Student: Sneha P Tiwari

Signature of Student:



Certificate

This is to certify that the Project work entitled Early Warning System using Global News for Investors and Creditors carried out by Sneha P Tiwari with SRN R19MBA08, is a bonafide student of REVA University, is submitting the project report in fulfilment of the award of PGDM in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score of less than 15%. The project work report has been approved as it satisfies the academic requirement in respect of PROJECT work prescribed for the said degree.



Akshay Kulkarni
Guide

Signature of the Director

Name: Dr. Shinu Abhi

External Viva
Names of the Examiners

1. <Name> <Designation> <Signature>
2. <Name> <Designation> <Signature>

Place: Bengaluru
Date:

Acknowledgment

The joy and satisfaction that follow any task's successful completion would be incomplete without considering the individuals who made it possible and under whose constant guidance and encouragement the task was completed.

I would like to express my immense gratitude to our Chancellor **Dr. P. Shyama Raju**, **Dr. S.Y. Kulkarni**, Ex-Vice Chancellor **Dr. D.K. Mallikharjuna Babu**, Vice-Chancellor **and Dr. Dhananjaya**, Registrar, for supporting the RACE program specifically designed for working professionals and providing facilities and infrastructure required and conducive conditions to offer the best learning experience. I am happy to be called as a part of this program and REVA university.

I would like to thank **Dr. Shinu Abhi**, Director, RACE (REVA Academics for Corporate Excellence), Bengaluru for her timely help and inspiration during the tenure of the Project work and course.

Sincerely I acknowledge our deep sense of gratitude to **Mr. Akshay Kulkarn**i, Lead Data Scientist, for his constant encouragement, help, and valuable guidelines.

I wish to extend my sincere thanks to all the **Mentors** of RACE, REVA, Bengaluru who has encouraged and guided us throughout the course.

At last but not the least, I express my heartfelt gratitude to the Almighty, my parents for their love and blessings that helped me to complete the Project Work successfully.

Place: Bengaluru
Date:

Name of the Student:
Sneha P Tiwari

Similarity Index Report

Title of the Thesis: Early Warning System using global news for Investors and Creditors

Total No. of Pages: 32

Name of the Student: Sneha P Tiwari

Name of the Guide(s): Akshay Kulkarni - Lead Data Scientist | Speaker | Author

This is to certify that the above thesis was scanned for the identification of similarities. The process and results are listed below.

Software Used: Turnitin
Date of Report Generation:
Similarity Index in %: 13%
Total word count: 5289

Place: Bengaluru
Date:

Name of the Student:
Sneha P Tiwari

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	FI	Financial Institution
2	GDELT	Global Database of Events, Languages, and Tone
3	GKG	Global Knowledge Graph
4	EWS	Early Warning System
5	GNPA	Gross Non-Performing Assets
6	SCB	Scheduled Commercial Banks

List of Figures

List of Tables

Abstract

In today's fast-changing world, an unforeseen scenario may pose a threat to a company's existence. This risk can affect a company's financial stability and damage its reputation in the market. And therefore, companies invest millions of dollars in Risk Management related activities. Risk management has now become an integral part of the corporate ecosystem. Risks for a company arrives wordlessly and has kept changing its forms may it be financial market volatility, project failure risks (at any point of life-cycle planning, growth, manufacturing or maintenance), legal liabilities, credit risk, accidents, disasters and natural causes, intentional attack from a competitor, or occurrences of unknown or unforeseen root-cause.

The ever-changing socio-economic landscape makes it extremely difficult to predict such emerging risks that cannot be simply fit into a grid. This creates a demand for a mechanism capable of reading the changing market trends, capturing hints of political/regulatory policy changes, quantifying the frequency and severity of human-made and natural disasters to help a company shield itself from external perils.

With the help of Data Science techniques, we can develop an early warning system in the form of a dashboard that can access the news media from both web and T.V. channels from across the world and accurately highlight the potential threats to the risk management team. Enabling the team to take all appropriate steps /precautions well in advance.

Keywords: EWS, Natural Language Processing, Rule-Based Monitoring

Contents

Chapter 1: Introduction

The primary foundation for any Investor or Financial institution is Risk management; it is the perfect embodiment of their portfolio applications and asset monitoring. Risk management's ultimate goal is to ensure that the loan fund is safe, profitable, and fluid. It is currently crucial for any financial institution to have an early warning system for bank risks. The financial crisis often causes such troubles, making the financial institution face a vast risk loss. Therefore, converting the bank credit risk measurement to measure enterprises' financial situation effectively reduces credit risk.

Enhancing risk management ability to respond to every new and unexpected financial situation and credit risk quantitative research is necessary for all financial institutions. A dire requirement to rethink the current methodologies and revise them as indicated by internationally acknowledged and progressed approaches utilized in developed economies. Financial Institutions need to move from an essential consistency driven post facto instrument to a proactive control-based advanced framework to screen such risks. From past experiences, we have learned that proactive controls help organizations restrict their exposure in bad segments and reduce overall losses.

To achieve this, financial organizations should continuously monitor all publicly available News, articles, etc. on borrower and industry-specific performances. The information gained from these sources must be analyzed on various analytical platforms, helping the lenders make timely decisions. Then the severity of the risk of the specific borrowers can be visualized in the form of a Dashboard; convenient mail can be generated based on the result of analysis and sent to the dedicated personnel of the risk analyst team, which will help support taking decisions in good time and also aids in focusing on the events that might be the source for upcoming risks.

Also, it might be noted that as credit quality-related data takes some time to affect the borrower, the framework can be intended to run through cluster measures, possibly day by day or week by week, instead of making it on an absolute real-time basis.

Chapter 2: Literature Review

The financial crisis has appeared in the emotional mould of why banks ought to screen credit risks. European Banks' credit risk cost has sky soared and continues to rise in certain markets. Even the banks those follow traditional, conservative, and once applauded approaches have endured. Banks often react by deleveraging when exposed to deteriorating credit portfolios. In April 2012, The International Monetary Fund alerted that further deleveraging might reduce the supply of credit and worsen customer's creditworthiness even more (McKinsey&Company, 2012). According to RBI data, on 31 August, considering the distribution of asset classes, the RBI estimated a rise in gross non-performing assets (GNPA) of Scheduled Commercial Banks (SCB) to 12.2% by March 2019 from 11.6% in March 2018. The primary reason for this being the inability to anticipate the incipient stress in portfolios that are likely to default (Pwc.in, 2019).

The stakeholders from investors and regulators anxiously want reassurances that banks know the skill of lending and are developing better risk-mitigations approach and credit monitoring systems (McKinsey&Company, 2012). Simultaneously, over the last two decades, the dynamically evolving regulations of RBI have reached an early warning identification approach from rule-based monitoring (from Income Recognition and Asset Classification (IRAC) norms in the Nineties to monitoring specific signals in 2017). Therefore, it is required to rework on the traditional practices by evolving into a proactive control-based digital system (Pwc.in, 2019). The financial institutions can meet this requirement by both early detection and effective mitigation of credit risks.

To achieve this, the publicly available news about the applicant- and industry-specific outcomes would have to be screened by the loan specialist. Analyzing this data on better statistical and analytical tool which would aid lenders to take opportune decisions (Pwc.in, 2019). One such system that attempts in predicting the risk of crisis using external environmental factors is called the Early Warning System (EWS). Research on comparable EWS showed that speculative attacks during political chaos, economic downturn, and inefficient regulatory environments can be caused by short-term debt and currency depreciation. The results also indicated that government instability, corruption, high short-term debt, unstable monetary, and fiscal policies do not only reduce investor's confidence but also

prevent effective crisis prevention strategies. Therefore, the financial institution will be able to monitor external environmental factors causing the crisis by adopting EWS (Tamadonejad et al., 2016).

Despite the extensive literature on the prediction of financial crises with the aid of using Early Warning Systems (EWSs), their realistic use even in the worldwide Financial Institutions by the policymakers is limited. This could be a contradiction due to the ever-changing nature of the financial risks as greater economies liberalize and expand their monetary systems. Along with the current ongoing innovations, making use of the outcomes of these EWS becomes more necessary than ever to prevent all kinds of financial crises (Davis & Karim, 2008).

In this context, an author in his study suggested that logit is the most appropriate approach for global EWS and signal extraction for country specific EWS. Besides, when designing predictive models and setting related thresholds it is important to consider the policy maker's objectives to correctly differentiate between actual crises and false alarms (Davis & Karim, 2008. Another study emphasis setting up an early warning indicator for alerting the credit risks for commercial banks in advance with the help of artificial intelligence. This approach is proved as an effective and objective method since it can provide a conceptual model, which is more reliable and scientific for identification and early warnings of the credit risk of commercial banks (ZhiYuan & ShuFang, 2013). The idea of tackling the design of an optimal EWS with two different angles was given by Fuertes the choice of the econometric methodology and the evaluation of the EWS itself. It compares a logit regression for credit ratings, a logit regression for macro data and K-means clustering of macro data, and the combined forecasts from all three methods (Fuertes & Kalotychou, 2007).

Chapter 3: Problem Statement

Recent financial emergency uncovered that numerous financial institutions (F.I.s) fizzled to recognize the rise in credit risk emerging due to ever-changing socio-economic landscapes in their portfolios early enough to require appropriate action. This is caused due to some of the major causes such as too much focus endorsing and compliance, need for a committed organizational unit, which is often called as EWS or Monitoring system, need enough intergroup and interdepartmental communication, lack of solutions which can be broader inclusive, faster, proactive and used to shield itself from external perils.

Currently, most of the Financial Risk Management teams determine the level of risk by using several Financial Risk Modelling Techniques. These models calculate the risks considering variables such as Balance sheets, Cash Flow Statements, economic conditions, etc. Few teams depend on external rating agencies such as Moody's, Fitch Ratings, S & P Global Ratings, etc. Credit MIS reporting, another method where the borrower will be asked to submit pre-determined financial statements to the lender periodically. These methods somehow miss incorporating risks involving various external factors such as financial markets, Legal liabilities, Natural causes and disasters, and many more. Hence FIs are looking forward to a solution to defend themselves from potential hazards.

An early warning system is a very fundamental unit to any financial institution exposed to monetary risks. It aids the F.I. to recognize early warning signals of an increase in risk soon enough and distinguish between clients whose default can be anticipated by taking fitting activities from those where a more aggressive procedural is optimal. In that sense, an Early warning system provides a competitive advantage over other investors, and it also aids in minimizing the losses related to investment risk. These early warning systems should also be able to provide warning signals considering not only the traditional risk modeling methods but also the other various external factors which are prone to cause risks in later stages, which would help the risk management team to make decisions based on the right time and right information.

Chapter 4: Objectives of the Study

An over-view target of this project is to revise the traditional Financial Risk Modeling methods by designing an Early Warning System that is capable of incorporating insights from various factors such as financial markets, Legal liabilities, Natural causes, and many other events in which the borrowers are directly or indirectly involved. The study focuses on extracting insights from the numerous News and articles published on both web and T.V. channels from across the world with the help of existing analytical tools and data science techniques. The early warning signals generated enable to take the appropriate precautions long in advance to prevent any sort of financial/reputational risk to the organization. Here are some of the benefits that a risk analyst can expect from this project.

1. The decrease in default rate at portfolio level:

   The default rate can be reduced at the portfolio level by restricting the focus to specific customers or industry where warning signals have been observed. This will help in controlling the probability of defaults.

2. Reinforcing of covenants:

   Lenders can intervene early by reinforcing the covenants and increase the collateral levels to achieve maximum recovery. Once a borrower account comes into the watch list, periodic and regular surveillance will aid in maximizing the collateral-related requirements and decrease losses caused due to the event of actual default.

3. Recognizing trends and opportunities in the market:

   New trends in various industries can be identified, and the organization can regulate its diverse portfolios. For instance, it may have helped banks identify the airline sector, turning adverse, and slowly reduce their exposure.

4. Revamping of the credit policies of organizations:

   Based on the information provided by the system about sectoral regulations, performance, and news, high-level credit policy-related decisions can be taken by the organization.

Chapter 5: Project Methodology

CRISP-DM methodology was used to execute this project structurally, where all the project activities were grouped into the following five predefined steps –

- Data understanding
- Data preparation
- Modeling,
- Evaluation
- Deployment

Figure No. 5.1 CRISP-DM Methodology

Each step mentioned above has been explained in detail in the following sections.

Extensive information on the problem statement and business case has been gathered by interviewing team personnel and colleagues from the Financial Risk Management team. After a few meetings and brainstorming sessions, an ideal format of the data and the required features to extract were decided. However, removing the desired data was difficult as the time window

selected was huge and could not compromise the amount of data to achieve a satisfactory outcome.

We had to choose one particular industry of interest during data collection to narrow down the scope. To select the industry initially, we had to extract sample data related across all major sectors and check for the quality of the data extracted, which was an iterative task. Another major challenge was observed when we bridged the gap between our obtained quantifiers and traditional Financial risk modeling variables.

The complete data operations from loading, cleaning, transforming, and modeling was done in Python. PowerBI is used to create an interactive dashboard to exhibit all the insights and quantifiers obtained after analyzing the transformed and modeled data.

Chapter 6: Business Understanding

The primary goal of any Financial Risk Analyst team is to have a system or a model which can be broader inclusive, faster, proactive that can be used to shield itself from external perils. With the current Risk Modelling Techniques which calculates risk rating based on the different features such as Balance sheets, Cash Flow Statements, economic conditions, etc. followed by the Financial Risk Analyst team, one fails to take into consideration the risks originating from various channels like unexpected events, uncertainty in financial markets, legal liabilities which also influences in damaging the company's reputation in the market along with the company's financial stability. The digital revolution that has transformed the ability to access the global News, articles, and such sources can update the present methodologies, which enables the system to include obtained insights and assist in making decisions with better coverage amongst the vast data by considering intuitions of public data and significant events across the globe.

Chapter 7: Data Understanding

The data utilized for the project is a part of GDELT or Global Database of Events, Language, and Tone (https://www.gdeltproject.org/, 2011), it is the first, open and most robust, database of human society ever developed. Creating a forum that tracks the world's news media in print, radio, and online formats, in over 100 languages, across the globe, each moment of the day, extending back to January 1, 1979, to the present day which updates every 15 minutes. It is created by Kalev Leetaru of Yahoo, Georgetown University, Philip Schrodt, and others, supported by Google Jigsaw. It is described as "an effort to create a catalogue of human social behaviour and beliefs across the globe, linking every person, organization, place, count, trend, news channel, incidents around the world into a huge network that captures what is happening around the world, what it means is and who is involved, and how the world feels about every day.

Today GDELT considers countless broadcasts, print, and online news sources from all over the world in not less than a hundred languages, and its source list develops every day. Notwithstanding overall interpreted news material, the verifiable backfile of GDELT extending back to 1979 utilizes The Washington Post, United Press International, The New York Times, Foreign Broadcast Information Service, Facts on File, Christian Science Monitor, BBC Monitoring, Associated Press Worldstream, AfricaNews, Agence France Presse, Associated Press, Associated Press Online(https://www.gdeltproject.org/, 2011).

The GDELT database itself has a different set of data such as GDELT Event Database, GDELT Global Knowledge Graph (GKG), GEDLT Visual GKG, and many more. The dataset we are using is GEDLT 2.0 GKG. The data is available in zip files in a tab-separated value format using a CSV extension making it easy to import into any spreadsheet software such as Microsoft Excel. The dataset can also be accessed using Google BigQuery. That is available on the Google Cloud Platform.

Chapter 8: Data Preparation

Understanding the vastness of GDELT data and considering the configurations, compatibility of our machine, we had to focus on a specific industry and reduce the time window, and to reduce our scope of the project.

Time window considered:	November 2019 to August 2020
Industry under the scope:	Automobile Industry
Companies of Interest:	Volkswagen, Hertz, Ford, Tesla, Scania, Renault, Nissan, and Toyota

Table No. 8.1

The data files for the given period and selected companies were extracted from the GDELT GKG 2.0 database using Python. While removing the data, it was made sure that the only required feature columns were selected. The below figure represents the sample of data extracted.

	DATE	DocumentIdentifier	V2Tone	Quotations
208035	2.020083e+13	https://borneobulletin.com.bn/2020/09/hamilton...	1.74129353233831,3.98009950248756,2.2388059701...	NaN
208036	2.020083e+13	https://www.autoblog.com/2020/08/31/2021-nissa...	0.766283524904215,2.68199233716475,1.915708812...	NaN
208037	2.020083e+13	https://www.autosport.com/f1/news/151803/10-th...	-1.43597925807738,2.63262863980854,4.068607897...	9479|27||certain members of the team#9702|59||...
208038	2.020083e+13	https://www.stuff.co.nz/motoring/122624646/toy...	3.39622641509434,4.15094339622642,0.7547169811...	NaN
208039	2.020083e+13	http://www.msn.com/en-nz/autos/news/analysis-w...	0.377928949357521,2.41874527588813,2.040816326...	2577|44||concern when there are global trade t...

Figure No. 8.1 Snippet of Raw Data

The feature columns selected among all columns of the GDELT GKG 2.0 dataset are as below.

- Date:

The data type of the column here in Integer. This is the Date in YYYYMMDDHHMMSS format on which the news media used to construct this GKG file was published. This Date is the publication date of the article from where the information was extracted

- DocumentIdentifier:

The data type of the column here is Text. It is the source document's unique external identifier. It can be used to identify the record and access it if the form is public and if you have all the required authorizations and subscriptions. This field consists of a series of values, from URLs

of open web resources to textual citations of print or broadcast material to DOI identifiers for various document repositories.

- V2Tone:

The data type of the column here is Floating Point. This field consists of a comma-delimited list of six score sentiment dimensions. Tone - It is the average "tone" of the overall article. This score lies between -100 (extremely negative) and +100 (extremely positive). Most general values range from -10 to +10, with 0 indicating neutral. Positive Score – It gives the percentage of all words in the article found to have a positive emotional implication. Lies between 0 - +100. Polarity- It gives the percentage of all words that had matches in the tonal dictionary to indicate how emotionally polarized or charged the Text is. Activity Reference Density and Self/Group Reference Density are two more dimensions that we have not considered for our scope of the study.

- Quotations:

Excerpted comments made by the participants in any incident are always featured by News media. Those influenced by it and these quotations will give critical input into the diverse viewpoints and thoughts around that incident. GDELT defines and extracts all cited phrases or comments from each of the source article.

To improve and ease V2Tone and Date columns' utilization, V2Tone columns have been split into three separate columns and converted the Date as string type. The below figure shows the snippet of transformed data.

	DocumentIdentifier	Quotations	Tone	Positive_Score	Negative_Score	Polarity	Date_N	Month_Year
208035	https://borneobulletin.com.bn/2020/09/hamilton...	NaN	1.741294	3.980100	2.238806	6.218905	20200831	202008
208036	https://www.autoblog.com/2020/08/31/2021-nissa...	NaN	0.766284	2.681992	1.915709	4.597701	20200831	202008
208037	https://www.autosport.com/f1/news/151803/10-th...	9479|27||certain members of the team#9702|59||...	-1.435979	2.632629	4.068608	6.701237	20200831	202008
208038	https://www.stuff.co.nz/motoring/122624646/toy...	NaN	3.396226	4.150943	0.754717	4.905660	20200831	202008
208039	http://www.msn.com/en-nz/autos/news/analysis-w...	2577|44||concern when there are global trade t...	0.377929	2.418745	2.040816	4.459562	20200831	202008

Figure No. 8.2 Snippet of Transformed Data

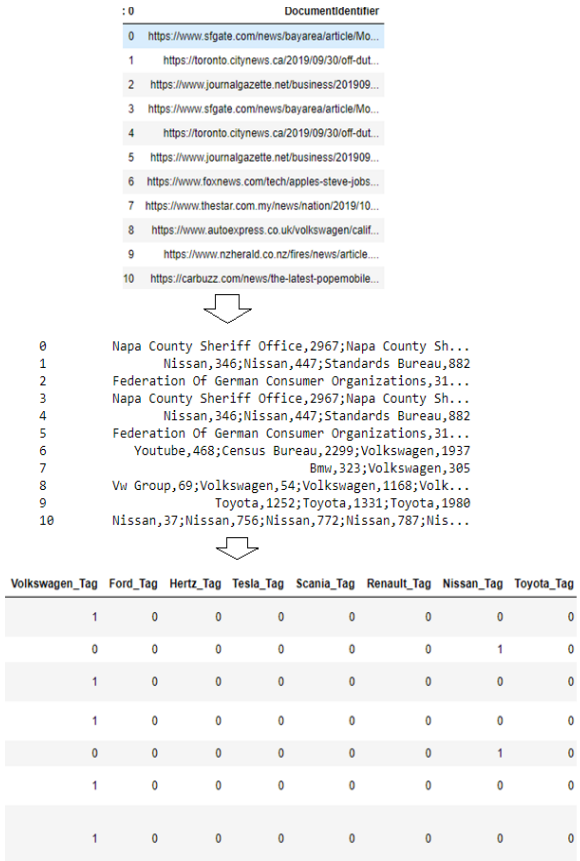
Chapter 9: Data Modeling

1. Web Scraping and Named Entity Recognition:

Once we had the URLs of all the News articles published about the desired companies for the given time window. The next task was to categorize the data into eight different companies of interest. To achieve the same below mentioned models and steps are made to run on the list URLs obtained from GKG 2.0 dataset.

- Web Scraping: All the content of a given URL is converted to a plain text using BeautifulSoup.
- Named Entity Recognition: Named Entity Recognition model is applied on the plain Text obtained after web scraping to identify all the organization the article is speaking about or mentioned.
- Further, label encoding is performed on the organizations' list for the given URL and is grouped under the eight specific companies.

Below snapshots show the different outputs of the models mentioned above and steps.



: 0	DocumentIdentifier
0	https://www.sfgate.com/news/bayarea/article/Mo...
1	https://toronto.citynews.ca/2019/09/30/off-dut...
2	https://www.journalgazette.net/business/201909...
3	https://www.sfgate.com/news/bayarea/article/Mo...
4	https://toronto.citynews.ca/2019/09/30/off-dut...
5	https://www.journalgazette.net/business/201909...
6	https://www.foxnews.com/tech/apples-steve-jobs...
7	https://www.thestar.com.my/news/nation/2019/10...
8	https://www.autoexpress.co.uk/volkswagen/calif...
9	https://www.nzherald.co.nz/fires/news/article....
10	https://carbuzz.com/news/the-latest-popemobile...

0	Napa County Sheriff Office,2967;Napa County Sh...
1	Nissan,346;Nissan,447;Standards Bureau,882
2	Federation Of German Consumer Organizations,31...
3	Napa County Sheriff Office,2967;Napa County Sh...
4	Nissan,346;Nissan,447;Standards Bureau,882
5	Federation Of German Consumer Organizations,31...
6	Youtube,468;Census Bureau,2299;Volkswagen,1937
7	Bmw,323;Volkswagen,305
8	Vw Group,69;Volkswagen,54;Volkswagen,1168;Volk...
9	Toyota,1252;Toyota,1331;Toyota,1980
10	Nissan,37;Nissan,756;Nissan,772;Nissan,787;Nis...

Volkswagen_Tag	Ford_Tag	Hertz_Tag	Tesla_Tag	Scania_Tag	Renault_Tag	Nissan_Tag	Toyota_Tag
1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0

Figure No. 9.1 Recognising and Grouping

2. WordCloud Generation:

Wordcloud is generated using the Text obtained from the Quotation column for all companies to know what perspectives and emotions the participant is putting in by finding out the frequently used words. Below are the steps used to generate the WordCloud.

- Missing rows are first eliminated since not all articles/ news published have excerpted statements. The below snapshot shows the sample set of Quotations.

	Quotations
11	2689|213||to approach various departments in the government towards fulfillment of conditions for our success in Trivandrum , which has been time and effort consuming , and other than running from...
22	3385|89||requested to pay over an amount of money into this WesBank Vehicle Finance Access Account#4454|41||successful fruit and vegetable wholesaler#10009|53||does not explain what the expectatio...
25	4248|56||haggled over the payment scheme with a new potential lab
30	2639|93||will be put on permanent display by the diocese as a memento of the pope visit to the island
39	2117|67||We are actually looking at a few initiatives that looks at warranty#2293|170||Let just say that we are working on a few things at the moment. It business case. If we can make it stack up ...
48	992|210||Autonomous vehicles will transform personal mobility x2026 ; xA0 ; reap the benefits of a new market which promises to ramp from essentially 0*nowto*10 trillion in global gross annual ...
53	2020|94||didn't exactly clarify the White House position , nor did it rule out other courses of action#3180|142||The Tokyo market is seen starting with falls , weighed down by losses in the US mar...
59	3003|33||Is this a genuine Ford accessory?
101	670|26||Timeless Japanese Futurism
110	3053|33||Is this a genuine Ford accessory?

Figure No. 9.2 Sample set of Quotations

- Later the Text is normalized by converting all the words to lower case and remove all unique character.
- NLTK is used to Stop word removal and perform stemming on the normalized Text to create a corpus.

```
['approach variou depart govern toward fulfil condit success trivandrum time effort consum run depart depart',
 'request pay amount money wesbank vehicl financ access account success fruit veget wholesal explain expect would',
 'haggl payment scheme new potenti lab',
 'put perman display dioces memento pope visit island',
 'actual look initi look warranti let say work thing moment busi case make stack come someth quit competit',
 'autonom vehicl transform person mobil x xa reap benefit new market promis ramp essenti trillion global gross annual reven
u',
 'exactli clarifi white hous posit rule cours action tokyo market seen start fall weigh loss us market investor eye chines ma
nufactur pmi like smaller previou tax hike',
 'genuin ford accessori',
 'timeless japanes futur',
 'genuin ford accessori',
 'want want',
 'electr motor deliv excel acceler high level respons effect stop go traffic even steep slope',
 'pass low qualiti materi inferior construct genuin part',
 'trade turn sluggish afternoon investor adopt wait see approach u china ministeri trade talk',
 'trade turn sluggish afternoon investor adopt wait see approach u china ministeri trade talk',
 'plan qashqai product sunderland chang sudden chang current arrang rule wto seriou implic british industri',
 'flow pattern mizuhiki thin twine made japanes rice paper',
```

Figure No. 9.3 Normalised Corpus

- Further, the corpus is converted to a string text file to create Wordcloud.

cleaned_text

' approach variou depart govern toward fulfil condit success trivandrum time effort consum run depart depart request pay amou nt money wesbank vehicl financ access account success fruit veget wholesal explain expect would haggl payment scheme new pote nti lab put perman display dioces memento pope visit island actual look initi look warranti let say work thing moment busi ca se make stack come someth quit competit autonom vehicl transform person mobil x xa reap benefit new market promis ramp essent i trillion global gross annual revenu exactli clarifi white hous posit rule cours action tokyo market seen start fall weigh l oss us market investor eye chines manufactur pmi like smaller previou tax hike genuin ford accessori timeless japanes futur g enuin ford accessori want want electr motor deliv excel acceler high level respons effect stop go traffic even steep slope pa ss low qualiti materi inferior construct genuin part trade turn sluggish afternoon investor adopt wait see approach u china m inisteri trade talk trade turn sluggish afternoon investor adopt wait see approach u china ministeri trade talk plan qashqai product sunderland chang sudden chang current arrang rule wto seriou implic british industri flow pattern mizuhiki thin twine made japanes rice paper brexit like sitcom start episod main charact tell casual lie abl skydiv impress someon end episod pla ne jump ask support brexit like ask punch constitu face said anna turley labour mp redcar vote leav make easier tell constitu want punch brexit like go back window may make amaz appl crumbl seismic shift trade condit seriou implic british industri ser iou implic british industri seriou implic british industri seriou implic british industri seriou implic british industri seri ou implic british industri seriou implic british industri seriou implic british industri seriou implic british industri colla bor research unam develop hand prosthesi softwar peopl think move extrem manag reduc thousand dollar thousand peso make acces s research center univers complic look research fund get support look submit five applic conacyt none pass project sent insti tuto nacion del emprendedor happen either research similar one propos conacyt accept gave dollar thank god peopl believ gave support seriou implic british industri seriou implic british industri seriou implic british industri seriou implic british in

Figure No. 9.4 Cleaned String Text

- Finally, WordCloud is generated from the string text file. Below snapshot represents the WordCloud developed for the company Nissan. The same will be generated for all the companies as part of the final Dashboard.

Figure No. 9.5 Generated WordCloud

3. Rule-Based model to Automate generation of warning emails to the team:

A simple condition-based model is created to automatically send out a warning mail to the team when the condition is satisfied.

This model is designed to compare the News published for a particular company on the latest Date with the Average frequency of the News published till Date. When the overall tone of the News published for the company on the latest Date is negative, a warning mail is generated and sent to the team.

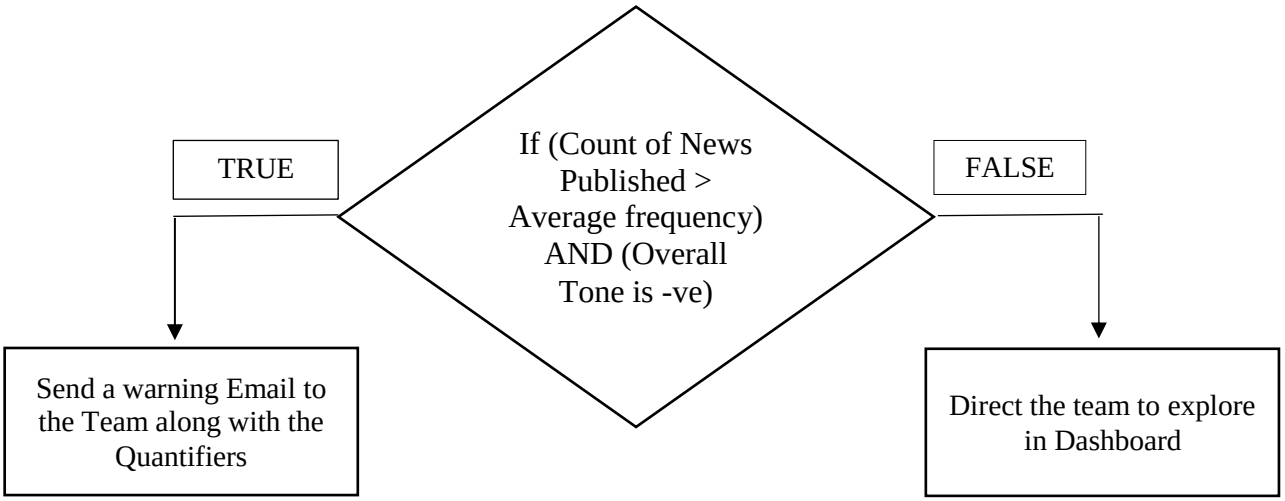


Figure No. 9.6 Rule to generate warning Email

Chapter 10: Data Evaluation

Many statistical methods and components were used to provide enough quantifiers and charts on the Dashboard to support the team to make better decisions and derive useful insights. Below are some of such helpful statistical components that we considered to include in the Dashboard.

1. Frequency of the news Published across the period:

This chart helps the analyst to understand how frequently the News was published about the company. And focus on the period where the frequency is higher, understand the cause for the same, and read the news article. Below is a trend chart generated using Python.

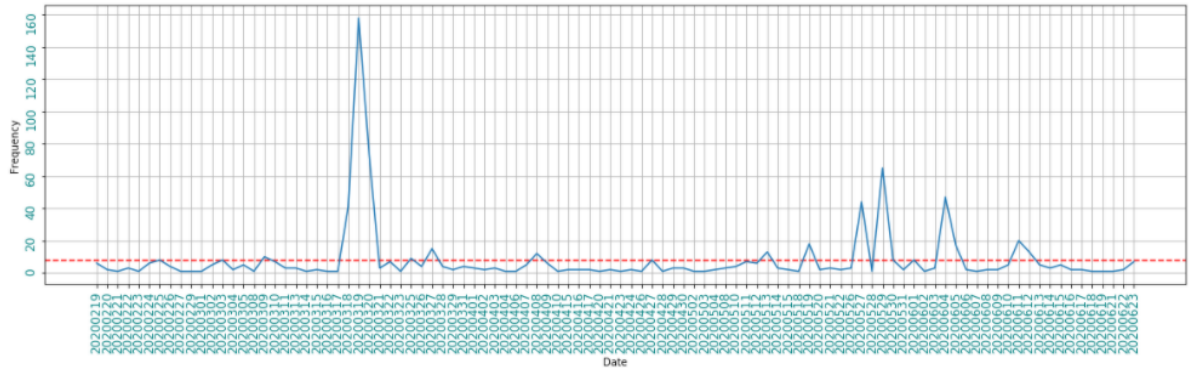


Figure No. 10.1 Frequency of the news published

2. The recency of the News Published:

This helps the analyst know when and how recently the News was published about the company, the list of recent news articles, and respective tonal scores to understand its sentiments better. Below once such a table is providing the current list of items and their score.

	Date_N	DocumentIdentifier	Positive_Score	Negative_Score	Polarity
208036	20200831	https://www.autoblog.com/2020/08/31/2021-nissan-frontier-interior-spy-photos/	2.681992	1.915709	4.597701
207744	20200831	https://europe.autonews.com/automakers/nissan-rebound-starts-next-year-ceo-says	3.623188	2.898551	6.521739
207755	20200831	https://www.trumbulltimes.com/news/article/Japan-s-Mitsubishi-executive-behind-Nissan-15526600.php	4.496788	4.496788	8.993576
207754	20200831	https://mynorthwest.com/2128642/japans-mitsubishi-executive-behind-nissan-alliance-has-died/	4.395604	4.395604	8.791209
207752	20200831	https://www.ctpost.com/news/article/Japan-s-Mitsubishi-executive-behind-Nissan-15526600.php	4.496788	4.496788	8.993576
207750	20200831	https://www.ctvnews.ca/world/japan-s-mitsubishi-executive-behind-nissan-alliance-has-died-1.5085720	4.075235	5.329154	9.404389
207748	20200831	http://bronx.news12.com/story/42566239/mitsubishi-motors-executive-behind-nissan-alliance-dies	4.444444	4.691358	9.135802
207746	20200831	https://www.msn.com/en-us/health/medical/love-your-neighbor-what-nashville-health-care-leaders-want-you-to-know-about-covid-19/ar-BB18xMch	2.761628	3.052326	5.813953
207741	20200831	https://www.wfmz.com/news/mitsubishi-motors-executive-behind-nissan-alliance-dies/article_e3b677bb-afec-5b97-bfeb-55819c498fad.html	4.736842	5.000000	9.736842
207813	20200831	https://www.autoblog.com/2020/08/31/mitsubishi-ceo-osamu-masuko-dies-71/	4.945055	5.219780	10.164835

Figure No. 10.2 List of Recent News Articles

3. Trend graph of the Sentiment Scores:

This component aids the analyst to understand the tone of the news articles published at the desired period and focus on any negative spikes in the company's trend.

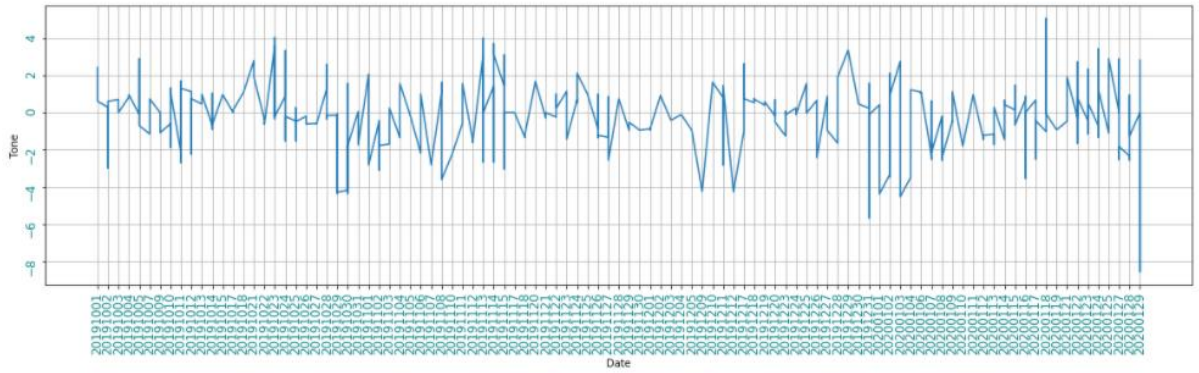


Figure No.10.3 Tone Trend Graph

4. Trend graph of the Polarity Scores:

Polarity is slightly different from the tone, which speaks about what emotion the author conveys and how much the tone indicator words match the author's feelings. This gives the analyst the right sentiment understanding of the articles.

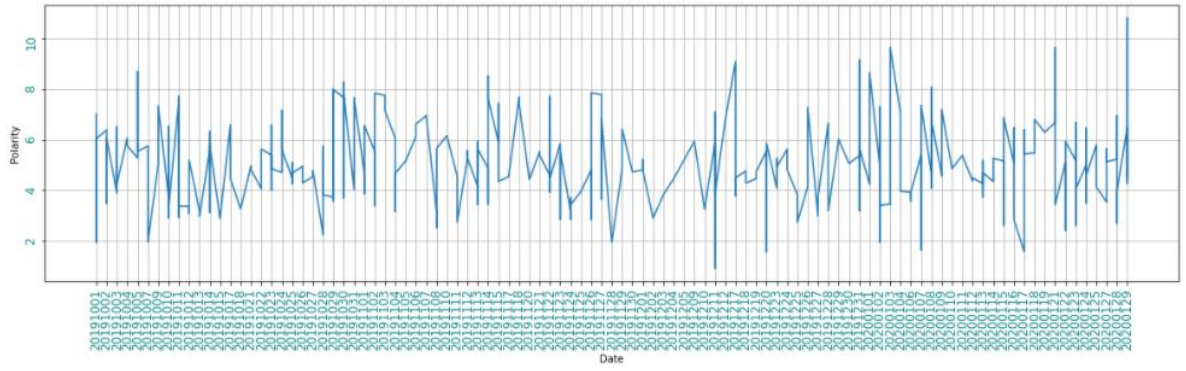


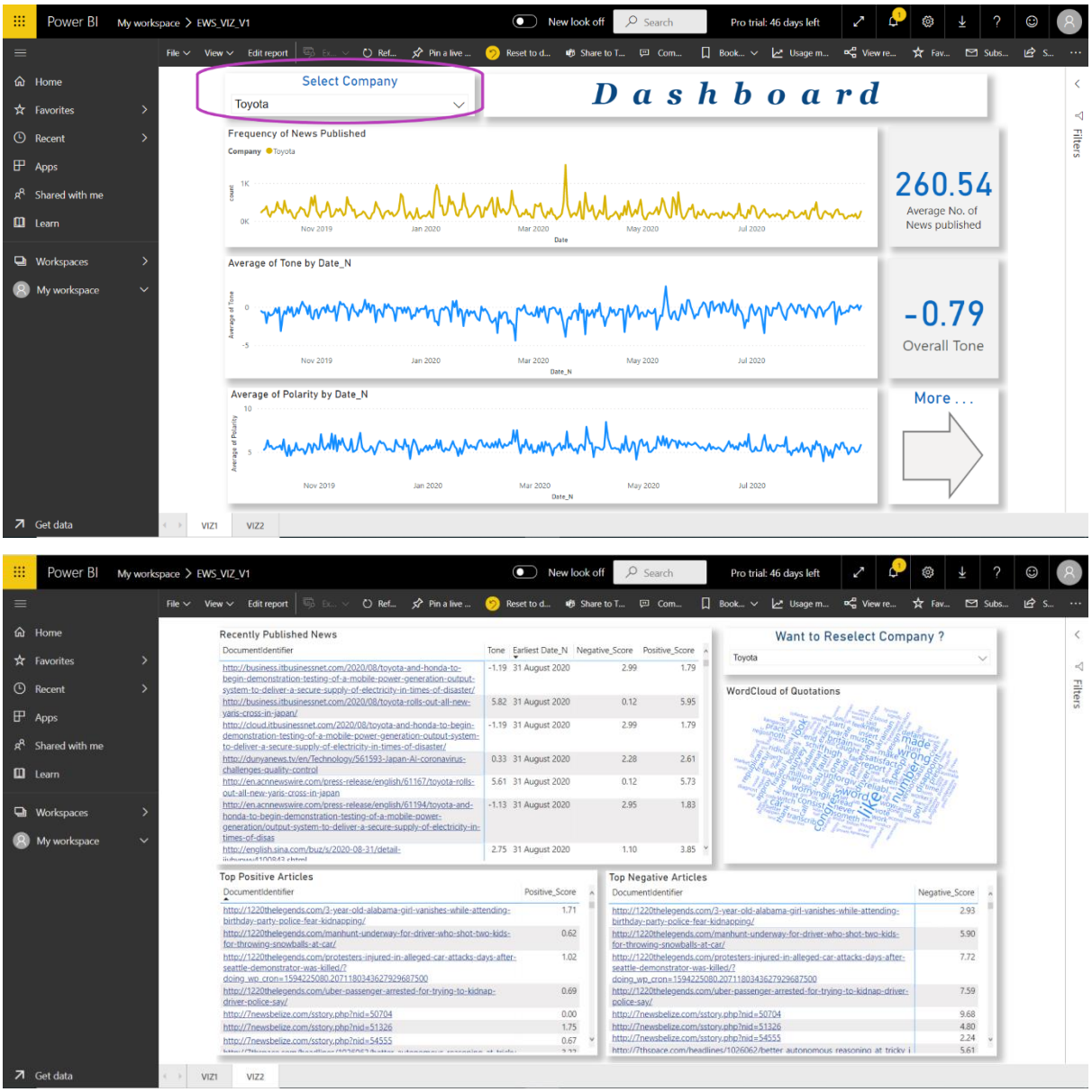
Figure No.10.4 Polarity Trend Graph

(Note: All the above-mentioned components are for a specific company of interest)

Chapter 11: Deployment

After deciding the comprehensive components and creating them, the same has been generated in PowerBI to create an interactive dashboard. The analyst can select the company of interest from the provided dropdown, and all the components for the company will be displayed. One can even click on the URLs and navigate to the article he wishes from the given list.

Below is the snapshot from the PowerBI Dashboard.



No.11.1 Snapshot of PowerBI Dashboard

Another part of the deployment is to create a test From Email ID and To Email ID to send and receive the warning Emails generated based on the condition. Below are snapshots of the warning Emails sent and received.

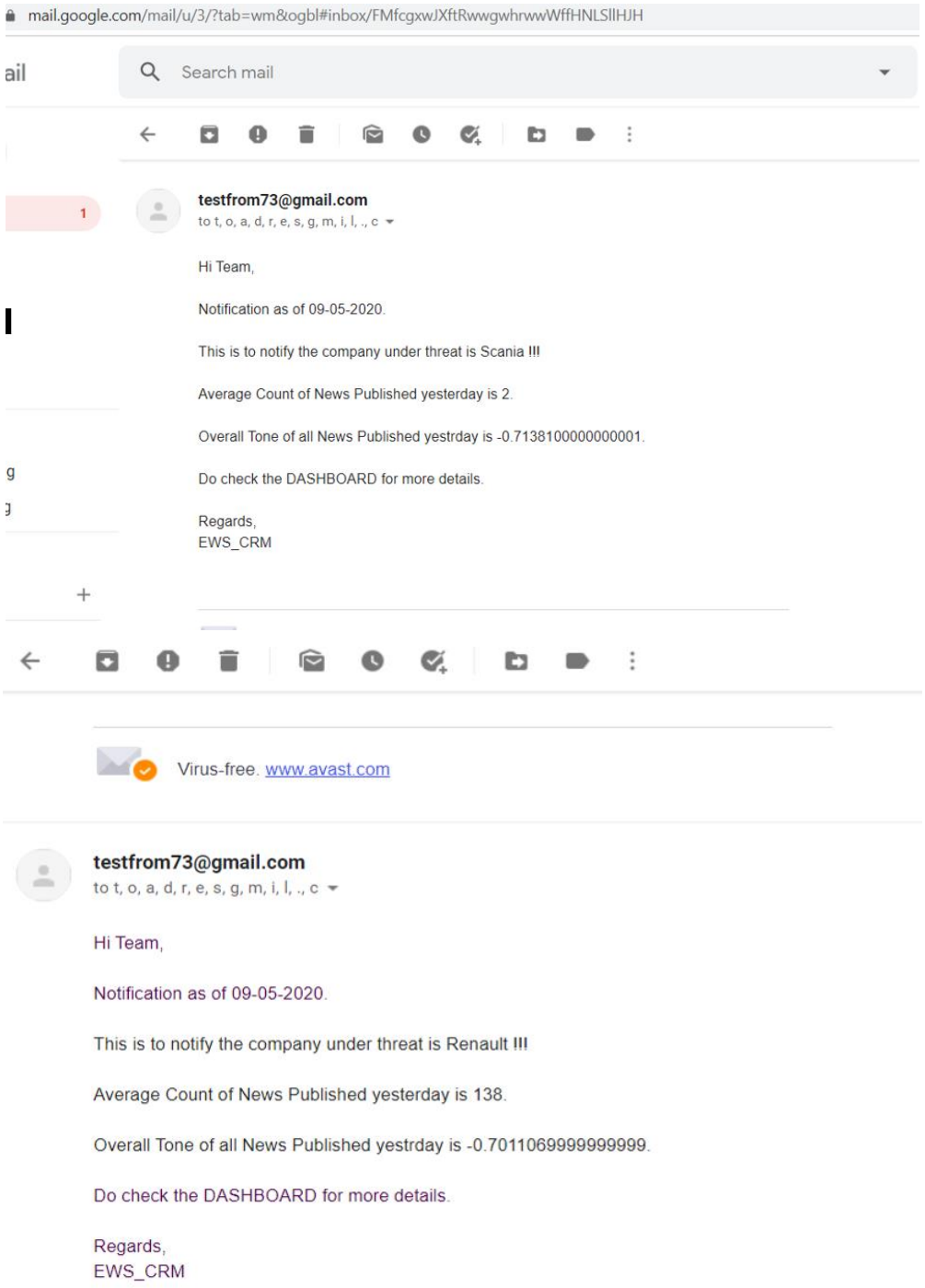


Figure No.11.2 Snapshot of Warning Emails received

Chapter 12: Conclusions and Recommendations for future work

As discussed, the overall goal of this project of creating a dashboard based Early Warning System was achieved by using GDELT GKG 2.0 data, which is a shared database of all publicly available articles, information. News published on all various sources such as Television, the Internet, and other local channels about the events/incidents happening worldwide. We have narrowed down our focus on the automobile industry by selecting and extracting all the required data from the aforementioned database for eight different automobile companies. Dashboard designed here would support a risk analyst team to make comprehensive, timely, and dynamic decisions so that they can take any preventive measures or reconsider any decisions made previously. This Dashboard provides sentiment scores of the news articles published about the company of interest; it contains a WordCloud that gives the analyst an essence of the article about the emotions and verbs used in the report's excerpted statements. An analyst is provided with a list of recent articles and their respective sentiment scores and can navigate to those articles and read if he wishes to; an analyst can choose his company of interest from the dropdown given and explore all the Dashboard mentioned above components. Also, few quantifiers are highlighted on the Dashboard, depending on which a rule-based model is created to send out automated generated warning Emails to the concerned team.

Considering the above-designed Dashboard and outputs, below are some of the recommendations for future work:

- Dynamic analysis and affinity analysis on the data extracted and discover any kind of relationships or insights on the activities, incidents, etc. recorded about the company of interest.

- Creating a network graph would help the analyst team focus on any other companies or organizations prone to be in the same bandwidth of risk as their company of interest.

- To create a similar dashboard with extensive components that works on the input given from the dropdown and enables the user to enter any keyword he wishes and gets all the possible insights from the available data about the keyword entered.

- The next major step would be to bridge an association between the Dashboard's output and the traditional credit risk modelling methods by converting the Dashboard outputs to a generalized rating system given to the company of interest.

Bibliography

-, Z. Y., & -, S. Z. (2013). Bank Credit Risk Management Early Warning and Decision-making based on BP Neural Networks. *INTERNATIONAL JOURNAL ON Advances in Information Sciences and Service Sciences*. https://doi.org/10.4156/aiss.vol5.issue9.51

Davis, E. P., & Karim, D. (2008). Comparing early warning systems for bankingDavis, E. P., & Karim, D. (2008). Comparing early warning systems for banking crises. Journal of Financial Stability. https://doi.org/10.1016/j.jfs.2007.12.004 crises. *Journal of Financial Stability*. https://doi.org/10.1016/j.jfs.2007.12.004

Fuertes, A. M., & Kalotychou, E. (2007). Optimal design of early warning systems for sovereign debt crises. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2006.07.001

https://www.gdeltproject.org/. (2011). *The GDELT Project*.

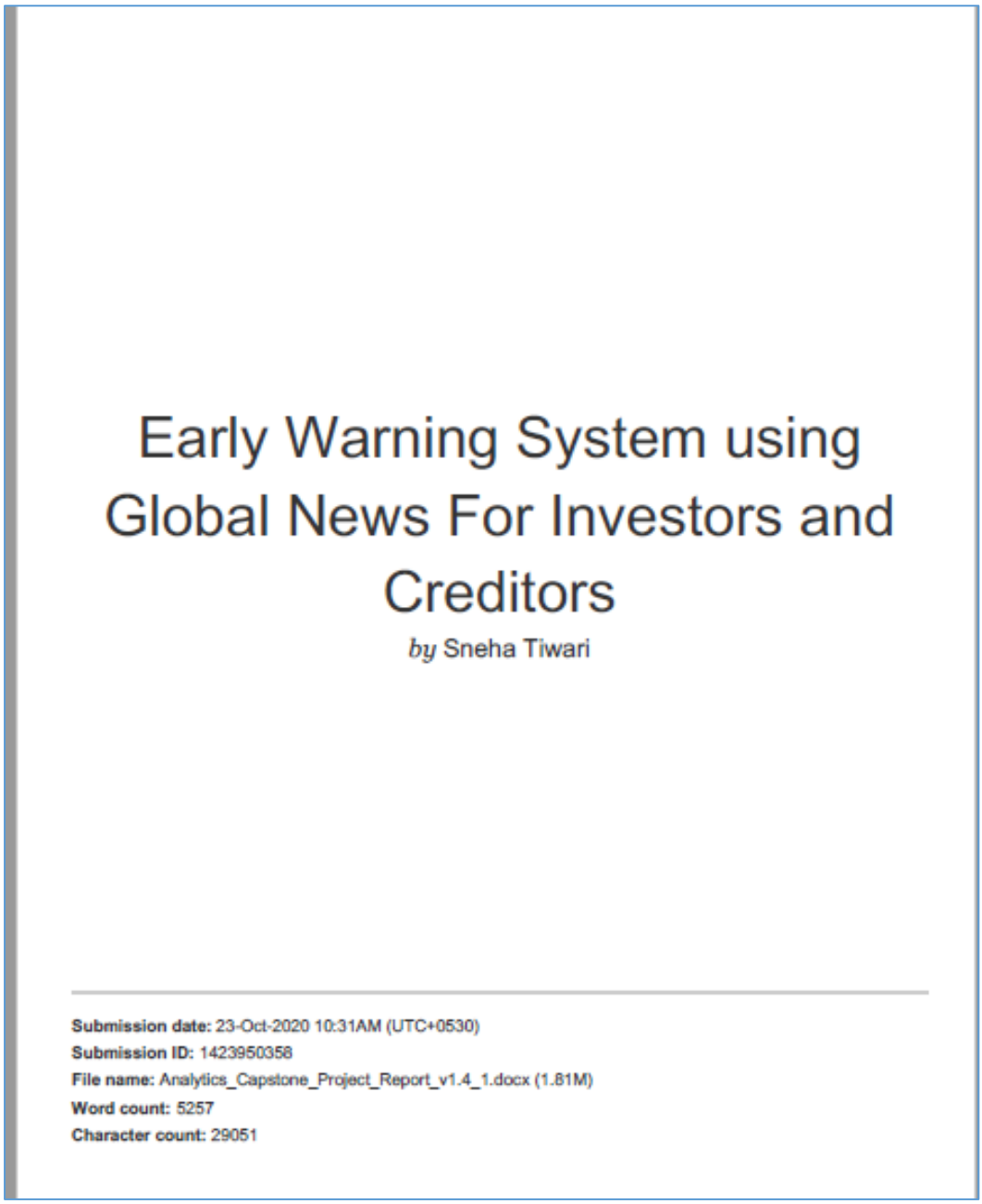
McKinsey&Company. (2012). Building credit monitoring for competitive advantage. *McKinsey Working Papers on Risk*. https://www.mckinsey.com/~/media/mckinsey/business functions/risk/our insights/a better way for banks to monitor credit/credit_monitoring_for_competitive_advantage(1).ashx

Pwc.in. (2019). *Early warning signals in a digital era: A proposed PwC framework to manage credit risk better*. https://www.pwc.in/assets/pdfs/consulting/financial-risk-and-regulations/early-warning-signals-in-a-digital-era.pdf

Tamadonejad, A., Abdul-Majid, M., Abdul-Rahman, A., & Jusoh, M. (2016). Early Warning systems for banking crises: Political and economic stability. *Jurnal Ekonomi Malaysia*. https://doi.org/10.17576/JEM-2016-5001-03

Appendix

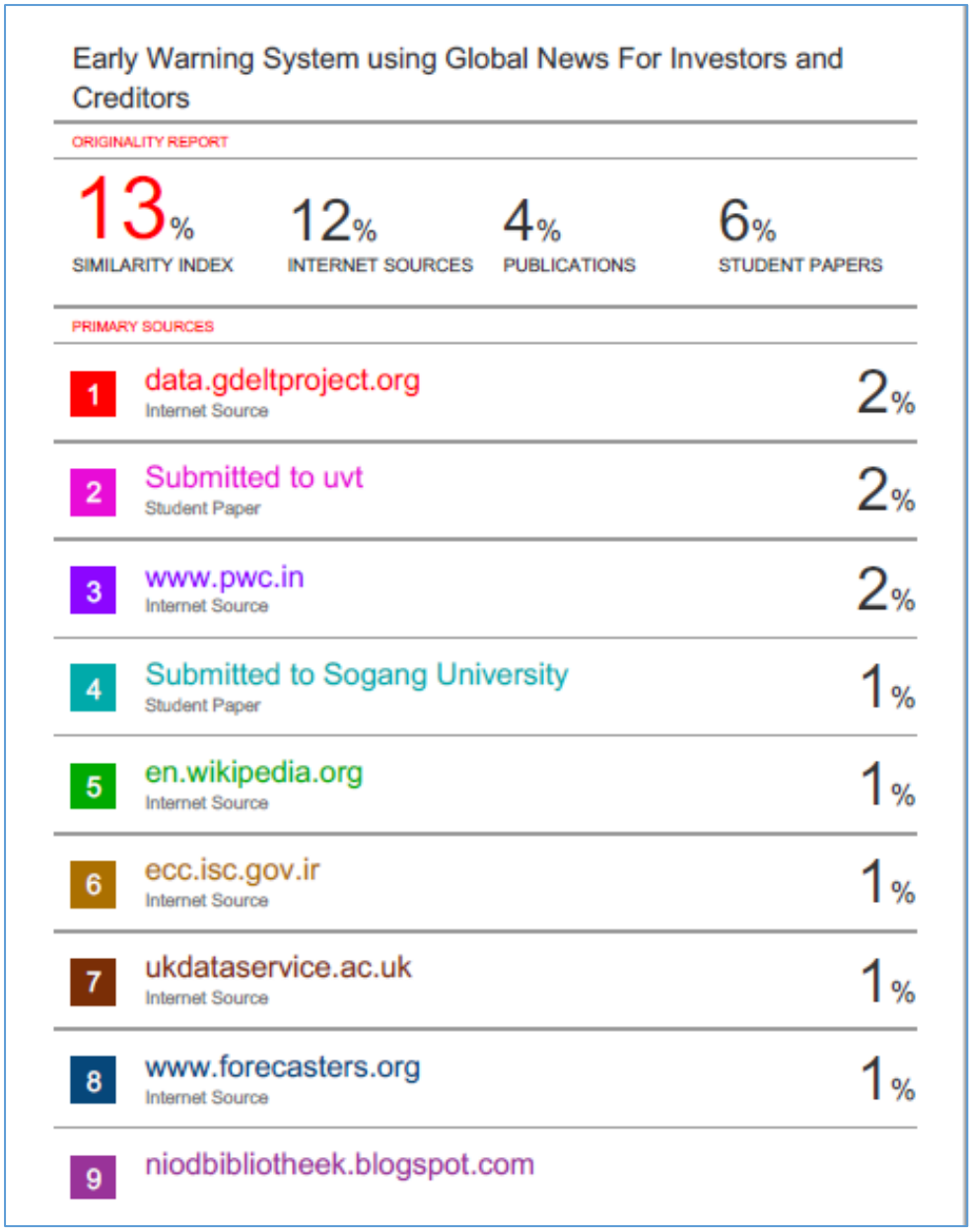
Plagiarism Report



Early Warning System using Global News For Investors and Creditors

by Sneha Tiwari

Submission date: 23-Oct-2020 10:31AM (UTC+0530)
Submission ID: 1423950358
File name: Analytics_Capstone_Project_Report_v1.4_1.docx (1.81M)
Word count: 5257
Character count: 29051



Early Warning System using Global News For Investors and Creditors

ORIGINALITY REPORT

13%	12%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	data.gdeltproject.org Internet Source	2%
2	Submitted to uvt Student Paper	2%
3	www.pwc.in Internet Source	2%
4	Submitted to Sogang University Student Paper	1%
5	en.wikipedia.org Internet Source	1%
6	ecc.isc.gov.ir Internet Source	1%
7	ukdataservice.ac.uk Internet Source	1%
8	www.forecasters.org Internet Source	1%
9	niodbibliotheek.blogspot.com	

	Internet Source	1%
10	Zhi-Yuan Yu, Shu-Fang Zhao. "Bank credit risk management early warning and decision-making based on BP neural networks", 2011 IEEE International Symposium on IT in Medicine and Education, 2011 Publication	1%
11	Submitted to University of Strathclyde Student Paper	1%
12	economice.ulbsibiu.ro Internet Source	<1%
13	www.scribd.com Internet Source	<1%
14	Submitted to University College London Student Paper	<1%
15	www.gdslink.com Internet Source	<1%
16	www.inmybangalore.com Internet Source	<1%
17	Yunpu Ma, Volker Tresp, Erik A. Daxberger. "Embedding models for episodic knowledge graphs", Journal of Web Semantics, 2019 Publication	<1%
	www.newsclick.in	