OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TECHNOLOGICAL UNIVERSITY DUBLIN

Text classification and clustering to detect antivaccine webpages

Dègninou Yehadji (B00108474)

Master of Science in Computing – Applied Data Science and Analytics

Technological University Dublin

Dublin 15

Assignment 2, Text Mining

I. Introduction

This project aims at performing classifications and clustering algorithms on a set of text documents. A combination of tools and software such as RapidMiner, R, and Excel were used. The cross-industry standard process for data mining (CRISP-DM) were used as methodological approach. Specifically, the following steps will be followed:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation

II. Business understanding

Vaccines have been proven effective in preventing a number of infectious diseases (Centers for Disease Control and Prevention (CDC), 2011). Yet, some people and organizations known as the antivaccine (antivax) movement have been spreading false information about vaccines. This movement has led to the increase of vaccine hesitancy in the population, and consequently, the reemergence of diseases that were considered as eradicated. The World Health Organization (WHO) has listed vaccine hesitancy as one of the threats to global health (WHO, 2019a), and major social media platforms are implementing strategies to detect and censor antivaccine information (Burki, 2019; Kari Paul and agencies, 2020; WHO, 2019b). Machine learning, specifically text mining is one of the approaches that can be used to detect vaccine misinformation. A text related to vaccine can be categorized as anti-vaccine (antivax), pro-vaccine (provax), neutral, or irrelevant. A text dataset of documents was collected from sources of these 4 categories to apply text mining approaches.

This project is intended to:

1) understand and visualize the documents in the dataset
2) design a model that can accurately build a classification model from a set of documents which would then be used to correctly assign test (unseen documents) to the correct class: provax, antivax, neutral, and irrelevant.

III. Data understanding

3.1. Data collection

The dataset was compiled from the sources presented in Table 1.

Table 1. *Sources of the dataset compiled for the project*

#	Source (website)	Class	Sample size	Dataset category	URL
1	Dr. Joseph Mercola	antivax	25	Training	https://articles.mercola.com
2	Vaccination Decisions	antivax	15	Test	http://vaccinationdecisions.net
3	Shot Of Prevention	provax	25	Training	https://shotofprevention.com
4	Adult Vaccines Now	provax	15	Test	https://adultvaccinesnow.org
5	BMC Infectious Diseases	neutral	25	Training	https://bmcinfectdis.biomedcentral.com
6	BMC Immunology	neutral	15	Test	https://bmcimmunol.biomedcentral.com
7	BMC Biochemistry	irrelevant	25	Training	https://bmcbiochem.biomedcentral.com
8	BMC Microbiology	irrelevant	15	Test	https://bmcmicrobiol.biomedcentral.com

The dataset contains a total of 160 documents weighting 2 to 44 Kb and split as follows: 100 documents for training the model and 60 documents to test the model.

3.2. Overview of the dataset

The dataset contains:

- 40 antivax documents, of which 25 are from Dr. Joseph Mercola website [https://articles.mercola.com] and used as training set, and 15 are from Vaccination Decisions [http://vaccinationdecisions.net] and used as test set (Table 2).
- 40 provax document, of which 25 are from Shot Of Prevention [https://shotofprevention.com] and used as training set, and 15 are from Adult Vaccines Now [https://adultvaccinesnow.org] and used as test set (Table 3).
- 40 neutral documents collected from two academic journals. These documents are classified as neutral because they are articles presenting basic or clinical research on vaccines, without strong statements promoting or discouraging vaccination. Only the abstracts of the articles were copied in documents used in this project. 25 of the documents are from BMC Infectious Diseases [https://bmcinfectdis.biomedcentral.com] and used as training set, and 15 are from BMC Immunology [https://bmcimmunol.biomedcentral.com] and used as test set (Table 4).
- 40 documents irrelevant to the vaccination topic. These documents are classified as irrelevant because they are extracted from academic journals publishing articles on several biological topics, except vaccines. Only the abstracts of the articles were copied in documents used in this project. 25 of the documents are from BMC Biochemistry [https://bmcbiochem.biomedcentral.com] and used as training set, and 15 are from BMC Microbiology [https://bmcmicrobiol.biomedcentral.com] and used as test set (Table 5).

All the 160 documents are webpages copied into text documents using Textise (CPC LLC, Austin, TX), an internet tool that removes images, forms, scripts, adverts, and leave plain text. The webpages' URLs were extracted from the websites using a Link Grabber.

Table 2. Antivax documents from Dr. Joseph Mercola and Vaccination Decisions websites

Source (website)	File	Dataset type	Document title
Dr. Joseph Mercola	antivax1	Training	Disabled Boy Kicked Out of School for Lack of Vaccination
Dr. Joseph Mercola	antivax2	Training	Flu Shot Fails to Protect Seniors and May Increase Miscarriages
Dr. Joseph Mercola	antivax3	Training	Vaccine Exemptions Under Attack in 2019
Dr. Joseph Mercola	antivax4	Training	The Disappearing Medical Exemption to Vaccination
Dr. Joseph Mercola	antivax5	Training	Chiropractor must pay $100k for sharing content like this
Dr. Joseph Mercola	antivax6	Training	Vaccine Injuries and Legal Mandates
Dr. Joseph Mercola	antivax7	Training	Trojan Horse of Measles — More Vaccines With the Mandate
Dr. Joseph Mercola	antivax8	Training	New mandates for hepatitis A vaccine
Dr. Joseph Mercola	antivax9	Training	Portions of Measles Outbreaks Are Due to Vaccine Reactions and Not Wild Measles Virus
Dr. Joseph Mercola	antivax10	Training	Vaxxed' — How Vaccine Safety Is Undermined and Suppressed
Dr. Joseph Mercola	antivax11	Training	Health Chief at NIH Denies Vaccine Injuries Occur
Dr. Joseph Mercola	antivax12	Training	US Vaccine Exemptions Remain Secure in 2018 — NVIC's 2018 Annual Report on US State Vaccine Legislation
Dr. Joseph Mercola	antivax13	Training	Why Herd Immunity Is a Hoax
Dr. Joseph Mercola	antivax14	Training	What is vaccine shedding?
Dr. Joseph Mercola	antivax15	Training	Moral Implications of Aborted Fetuses in Vaccine Production
Dr. Joseph Mercola	antivax16	Training	Nearly half of those surveyed doubt vaccine safety
Dr. Joseph Mercola	antivax17	Training	Should Pediatricians 'Fire' Patients Whose Parents Don't Vaccinate?
Dr. Joseph Mercola	antivax18	Training	CDC: HPV, the 'Anti-Cancer' Vaccine?
Dr. Joseph Mercola	antivax19	Training	Now More Vaccines Recommended for Failing Mumps Vaccine
Dr. Joseph Mercola	antivax20	Training	Flu Vaccination: The Hidden Risks in This Heavily Promoted Seasonal Routine
Dr. Joseph Mercola	antivax21	Training	Vaccine Court Is a Big Pharma Fraud
Dr. Joseph Mercola	antivax22	Training	One of the Most Inexcusable Vaccine Revelations of All
Dr. Joseph Mercola	antivax23	Training	Flu Vaccine Exposed
Dr. Joseph Mercola	antivax24	Training	Vaccine Can Wreck Your Immune System
Dr. Joseph Mercola	antivax25	Training	Eating Refined Sugar Increases Your Risk of Getting a Brain-Invading Virus
Vaccination Decisions	antivax26	Test	Australian Journalist, Cathy O'Leary, Promoting HPV Vaccine (Cervical Cancer) with Lies
Vaccination Decisions	antivax27	Test	Australian Media Providing False Information on PhD Vaccination Research
Vaccination Decisions	antivax28	Test	Coercive Measures in the Australian Government's Vaccination Policies
Vaccination Decisions	antivax29	Test	Conflicts of Interest in Government Vaccination Policies
Vaccination Decisions	antivax30	Test	Corruption in the NSW Health Department and HCCC
Vaccination Decisions	antivax31	Test	Do you know what's in a vaccine?
Vaccination Decisions	antivax32	Test	Dr. James Lyons-Weiler, Public Health Expert interviews Dr. Judy Wilyman
Vaccination Decisions	antivax33	Test	Vaccine Myths Promoted by the Australian Government
Vaccination Decisions	antivax34	Test	A Lack of Integrity in Australian Journalism and Government Vaccination Policy
Vaccination Decisions	antivax35	Test	A National Conference and Protest Against Australia's Coercive Vaccination Policies
Vaccination Decisions	antivax36	Test	WA Premier, Mark McGowan, tells Parents of Vaccine-Injured Children to "Bugger Off"
Vaccination Decisions	antivax37	Test	Mainstream Media will not Present our Position of Choice in Vaccination
Vaccination Decisions	antivax38	Test	Richard Di Natale's (Greens Leader) False Comments on Vaccination made in the Australian Parliament

Vaccination Decisions	antivax39	Test	Introducing a 'No Jab No Play' Policy into Western Australia
Vaccination Decisions	antivax40	Test	Australian Journalist, Cathy O'Leary, Promoting HPV Vaccine (Cervical Cancer) with Lies

Table 3. Provax documents from Shot Of Prevention and Adult Vaccines Now websites

Source (website)	File	Dataset type	Document title
Shot Of Prevention	provax1	Training	Protecting Myself and My Child against Vaccine-Preventable Diseases during Pregnancy
Shot Of Prevention	provax2	Training	With School Vaccine Exemptions on the Rise, What Can Be Done to Protect Our Students?
Shot Of Prevention	provax3	Training	Rise in Vaccine Hesitancy Related to Pursuit of Purity: A Conversation with Professor Larson
Shot Of Prevention	provax4	Training	Five Important Reasons to Vaccinate Your Child
Shot Of Prevention	provax5	Training	Shingles Vaccine is the Silver Lining of Turning 50
Shot Of Prevention	provax6	Training	How Flu Strains are Selected for the Seasonal Flu Vaccine Each Year
Shot Of Prevention	provax7	Training	The State of the ImmUnion: A Report on Vaccine-Preventable Diseases in the U.S.
Shot Of Prevention	provax8	Training	Free Online Course Explains Cells, Immunity and Vaccines
Shot Of Prevention	provax9	Training	Flu Vaccine Benefits Go Beyond Effectiveness of One Strain
Shot Of Prevention	provax10	Training	Five Things I've Learned About Vaccines Through 21 Years of Parenting
Shot Of Prevention	provax11	Training	March for Science, Chant for Vaccines
Shot Of Prevention	provax12	Training	Federal & State Legislators are Listening: Time To Advocate For the Value of Vaccines
Shot Of Prevention	provax13	Training	Autism Expert Agrees: It's Time to Shift the Focus Off of Vaccines
Shot Of Prevention	provax14	Training	Sorry Kennedy, Being Anti-Vaccine Does Not Mean You're Pro-Safe Vaccine
Shot Of Prevention	provax15	Training	How Fake Vaccine News Is Dangerous to Us All
Shot Of Prevention	provax16	Training	Multiple Vaccine Oversight Committees Ensure Our Public Safety
Shot Of Prevention	provax17	Training	One Mom's Advice: Get Your Kids A Flu Vaccine As Soon As It Becomes Available
Shot Of Prevention	provax18	Training	Your Amazon Purchases Can Help Educate People About Vaccines
Shot Of Prevention	provax19	Training	Preventing Shingles Today and the Promise of New Vaccines Tomorrow
Shot Of Prevention	provax20	Training	Understanding Why Your Baby Needs a Hepatitis B Vaccine At Birth
Shot Of Prevention	provax21	Training	My Polio Story is an Inconvenient Truth to Those Who Refuse Vaccines
Shot Of Prevention	provax22	Training	Don't Just Vaccinate Your Kids, Teach Them the Science Behind Vaccines
Shot Of Prevention	provax23	Training	You Could be One Vaccination Away from Preventing Cancer
Shot Of Prevention	provax24	Training	Kids Team Up With Vaccine Heroes to Destroy Deadly Germs
Shot Of Prevention	provax25	Training	Parents Play Key Role as Local & National Vaccine Advocates
Adult Vaccines Now	provax26	Test	Dara Lieberman, MPP, Trust for America's Health
Adult Vaccines Now	provax27	Test	AVAC Blog Spotlight: Amy Pisani, Vaccinate Your Family
Adult Vaccines Now	provax28	Test	Healthy People 2030
Adult Vaccines Now	provax29	Test	Building a Strong Vaccine Infrastructure
Adult Vaccines Now	provax30	Test	AVAC Spotlight: Serese Marotta, Families Fighting Flu (FFF)
Adult Vaccines Now	provax31	Test	AVAC Spotlight: Robert H. Hopkins, Jr., MD,
Adult Vaccines Now	provax32	Test	AVAC Spotlight: Litjen (L.J.) Tan, MS, PhD, Chief Strategy Officer of Immunization Action Coalition
Adult Vaccines Now	provax33	Test	AVAC Spotlight: Dr. Wilbur H. Chen
Adult Vaccines Now	provax34	Test	Measles Madness And Value: Gene Therapy, Prevention, And The Pre-Vaccine Era
Adult Vaccines Now	provax35	Test	Vaccinate Your Family Mourns the Loss of Rich Greenaway
Adult Vaccines Now	provax36	Test	Moving More Electrons To Optimize New Adult Composite Immunization Measures
Adult Vaccines Now	provax37	Test	AVAC Spotlight: Lauren Linkenauger, PharmD
Adult Vaccines Now	provax38	Test	AVAC Spotlight: Michael Popovich, CEO, Scientific Technologies Corporation (STC)

Adult Vaccines Now	provax39	Test	AVAC Spotlight: Sarah Irsik-Good, MHA, President and CEO of the Kansas Foundation …
Adult Vaccines Now	provax40	Test	Promoting Access Starts With Educating People About Need

Table 4. Neutral documents from BMC Infectious Diseases and BMC Immunology journals

Source (website)	File	Dataset type	Document title
BMC Infectious Diseases	neutral1	Training	Evaluation of non-inferiority of intradermal versus adjuvanted …
BMC Infectious Diseases	neutral2	Training	The immunogenicity and safety of a reduced PRP-content DTPw-HBV/Hib vaccine …
BMC Infectious Diseases	neutral3	Training	A phase II, open-label, multicentre study to evaluate the immunogenicity and …
BMC Infectious Diseases	neutral4	Training	Dynamic models of pneumococcal carriage and the impact of the Heptavalent …
BMC Infectious Diseases	neutral5	Training	Why do I need it? I am not at risk! Public perceptions towards the pandemic (H1N1) 2009 vaccine
BMC Infectious Diseases	neutral6	Training	Effectiveness of a MF-59™-adjuvanted pandemic influenza vaccine to prevent…
BMC Infectious Diseases	neutral7	Training	An analysis of national target groups for monovalent 2009 pandemic influenza …
BMC Infectious Diseases	neutral8	Training	Flexibility of interval between vaccinations with AS03A-adjuvanted influenza …
BMC Infectious Diseases	neutral9	Training	A randomized, controlled non-inferiority trial comparing A(H1N1)pmd09 vaccine …
BMC Infectious Diseases	neutral10	Training	Usefulness of health registries when estimating vaccine effectiveness during the …
BMC Infectious Diseases	neutral11	Training	HLA-A*0201-specific epitopes of Indian HIV-1C as candidates for vaccine design
BMC Infectious Diseases	neutral12	Training	Development of a vaccine delivery system using hepatitis B core antigen based…
BMC Infectious Diseases	neutral13	Training	Immune response, antibody persistence, and safety of a single dose of the …
BMC Infectious Diseases	neutral14	Training	Immunogenicity and safety of quadrivalent versus trivalent inactivated influenza …
BMC Infectious Diseases	neutral15	Training	Immunogenicity, reactogenicity and safety of an inactivated quadrivalent …
BMC Infectious Diseases	neutral16	Training	Selection of an adjuvant for seasonal influenza vaccine in elderly people…
BMC Infectious Diseases	neutral17	Training	A historically-controlled Phase III study in adults to characterize the acceptability …
BMC Infectious Diseases	neutral18	Training	An adjuvanted inactivated murine cytomegalovirus (MCMV) vaccine induces …
BMC Infectious Diseases	neutral19	Training	A model-based analysis: what potential could there be for a S. aureus vaccine …
BMC Infectious Diseases	neutral20	Training	Superior antigen-specific CD4+T-cell response with AS03-adjuvantation of a…
BMC Infectious Diseases	neutral21	Training	A randomised trial to evaluate the immunogenicity, reactogenicity, and safety …
BMC Infectious Diseases	neutral22	Training	Long-term protection of hepatitis B vaccine in HIV-infected patients
BMC Infectious Diseases	neutral23	Training	In vivo analysis of GenePro, a lentivral therapeutic vaccine
BMC Infectious Diseases	neutral24	Training	Optimized production of a safe and efficient gene therapeutic vaccine …
BMC Infectious Diseases	neutral25	Training	Lentiviral-based anti-HIV therapeutic vaccine: design, preclinical studies…
BMC Immunology	neutral26	Test	Responses to pandemic ASO3-adjuvanted A/California/07/09 H1N1 influenza…
BMC Immunology	neutral27	Test	Tumor vaccine composed of C-class CpG oligodeoxynucleotides and irradiated …
BMC Immunology	neutral28	Test	Adjuvant effect of docetaxel on the immune responses to influenza A H1N1 vaccine in mice
BMC Immunology	neutral29	Test	Anti-idiotypic antibodies: a new approach in prion research
BMC Immunology	neutral30	Test	Enhancement of the priming efficacy of DNA vaccines encoding dendritic …
BMC Immunology	neutral31	Test	Using epitope predictions to evaluate efficacy and population coverage …
BMC Immunology	neutral32	Test	Adenovirus F protein as a delivery vehicle for botulinum B
BMC Immunology	neutral33	Test	Associations between SNPs in candidate immune-relevant genes and …
BMC Immunology	neutral34	Test	In vivo trafficking and immunostimulatory potential of an intranasally …
BMC Immunology	neutral35	Test	Intranasal immunization with plasmid DNA encoding spike protein …
BMC Immunology	neutral36	Test	Host immunity in the protective response to nasal immunization with …
BMC Immunology	neutral37	Test	Ontology-based Brucella vaccine literature indexing and systematic analysis …
BMC Immunology	neutral38	Test	Immune enhancement by novel vaccine adjuvants in autoimmune-prone …

BMC Immunology	neutral39	Test	Neuroantigen-specific, tolerogenic vaccines: GM-CSF is a fusion partner …
BMC Immunology	neutral40	Test	Comparative immunological evaluation of recombinant Salmonella Typhi…

Table 5. Irrelevant documents from BMC Biochemistry and BMC Microbiology websites

Source (website)	File	Dataset type	Document title
BMC Biochemistry	irrelevant1	Training	Exploring the functional interaction between POSH and ALIX and the relevance…
BMC Biochemistry	irrelevant2	Training	Mapping of protein phosphatase-6 association with its SAPS domain regulatory…
BMC Biochemistry	irrelevant3	Training	Mapping of protein phosphatase-6 association with its SAPS domain regulatory…
BMC Biochemistry	irrelevant4	Training	Development of a sensitive non-radioactive protein kinase assay and its appli…
BMC Biochemistry	irrelevant5	Training	Molecular characterization of tlyA gene product, Rv1694 of Mycobacterium tube…
BMC Biochemistry	irrelevant6	Training	High affinity binding of hydrophobic and autoantigenic regions of proinsulin…
BMC Biochemistry	irrelevant7	Training	Streptavidin-Binding Peptide (SBP)-tagged SMC2 allows single-step affinity fl…
BMC Biochemistry	irrelevant8	Training	Glycoproteomic characterization of carriers of the CD15/Lewisx epitope on Hod…
BMC Biochemistry	irrelevant9	Training	Elucidating the domain architecture and functions of non-core RAG1: The capac…
BMC Biochemistry	irrelevant10	Training	Signal peptide cleavage is essential for surface expression of a regulatory T…
BMC Biochemistry	irrelevant11	Training	Matrix metalloproteinase-19 inhibits growth of endothelial cells by generatin…
BMC Biochemistry	irrelevant12	Training	Gallus gallus NEU3 sialidase as model to study protein evolution mechanism ba…
BMC Biochemistry	irrelevant13	Training	The major leucyl aminopeptidase of Trypanosoma cruzi (LAPTc) assembles into a…
BMC Biochemistry	irrelevant14	Training	Application of Celluspots peptide arrays for the analysis of the binding spec…
BMC Biochemistry	irrelevant15	Training	Three genes expressing Kunitz domains in the epididymis are related to genes…
BMC Biochemistry	irrelevant16	Training	Topological characterisation and identification of critical domains within gl…
BMC Biochemistry	irrelevant17	Training	Myosin-cross-reactive antigen (MCRA) protein from Bifidobacterium breve is a…
BMC Biochemistry	irrelevant18	Training	Quaternary structures of recombinant, cellular, and serum forms of Thymidine…
BMC Biochemistry	irrelevant19	Training	Identification of critical residues of the serotype modifying O-acetyltransfe…
BMC Biochemistry	irrelevant20	Training	In vitro substrate phosphorylation by Ca2+/calmodulin-dependent protein kinas…
BMC Biochemistry	irrelevant21	Training	Ski-interacting protein (SKIP) interacts with androgen receptor in the nucleu…
BMC Biochemistry	irrelevant22	Training	Transcription start sites and epigenetic analysis of the HSD17B10 proximal pr…
BMC Biochemistry	irrelevant23	Training	Identification of the lamin A/C phosphoepitope recognized by the antibody P-S…
BMC Biochemistry	irrelevant24	Training	Proteomic analysis of differentially expressed proteins in vitamin C-treated…
BMC Biochemistry	irrelevant25	Training	KCTD20, a relative of BTBD10, is a positive regulator of Akt…
BMC Microbiology	irrelevant26	Test	Mycobacterium tuberculosis Rv0679c protein sequences involved in host-cell in…
BMC Microbiology	irrelevant27	Test	Persistent, triple-virus co-infections in mosquito cells…
BMC Microbiology	irrelevant28	Test	Comparison of BCG, MPL and cationic liposome adjuvant systems in leishmanial…
BMC Microbiology	irrelevant29	Test	Identification, expression and serological evaluation of the recombinant ATP…
BMC Microbiology	irrelevant30	Test	Genetic analysis of the capsule polysaccharide (K antigen) and exopolysacchar…
BMC Microbiology	irrelevant31	Test	Development of O-antigen gene cluster-specific PCRs for rapid typing six epid…
BMC Microbiology	irrelevant32	Test	Characterization of 13 multi-drug resistant Salmonella serovars from differen…
BMC Microbiology	irrelevant33	Test	Identification of a human immunodominant T-cell epitope of mycobacterium tube…
BMC Microbiology	irrelevant34	Test	Development of an indirect competitive enzyme-linked immunosorbent assay appl…
BMC Microbiology	irrelevant35	Test	Genesis of a novel Shigella flexneri serotype by sequential infection of sero…
BMC Microbiology	irrelevant36	Test	Immunoproteomics based identification of thioredoxin reductase GliT and novel…
BMC Microbiology	irrelevant37	Test	Macropinocytosis is responsible for the uptake of pathogenic and non-pathogen…
BMC Microbiology	irrelevant38	Test	Detection of Burkholderia pseudomallei O-antigen serotypes in near-neighbor s…
BMC Microbiology	irrelevant39	Test	An extracellular Staphylococcus epidermidis polysaccharide: relation to Polys…

3.3. Document vector indexing

Vector space model is a text document representation model used in information retrieval, indexing, information filtering, and relevancy rankings, in form of vectors of identifiers, such as index terms. In this model, documents are represented as vectors where each dimension represents a separate term that takes a non-zero value when the term is present in the document. Vector coefficients or lengths or term weights represent terms presence or importance.

There are several metrics to perform algebraic representation of vector space models:

- Binary: 0 when term is absent, 1 when term is present in the document;
- Thresholding frequencies to three values: 0 when term is absent in the document, 1 when term occurs once in the document, 2 when term occurs more than once in the document;
- Term occurrence: how often the term appears in the document;
- Term frequencies (TF): how often the term appears in the document divided by the number of terms in the document;
- Inverse document frequency (IDF): common terms are weighted close to 0 while fewer common terms are weighted close to 1;
- Term frequency-inverse document frequency (TF-IDF): a combination of local and global dictionary information;
- Latent Semantic Indexing: terms from related documents (Gudivada et al., 2018).

TF-IDF is the metric used for document vector indexing in this project. TF-IDF was computed, but only the values for the first documents of each class have been presented in this report (Table 6 and Table 7).

The training documents were fed into the Process Documents from Files operator by class and preprocessed without and after pruning (prune below 3% and above 30%). Transform case (lower case), Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English), and Filter Token by POS Tags (JJ.*|N.*|R.*|V.*) operators were used to preprocess the documents inside Process Documents from Files operator. The output was saved into an Excel file.

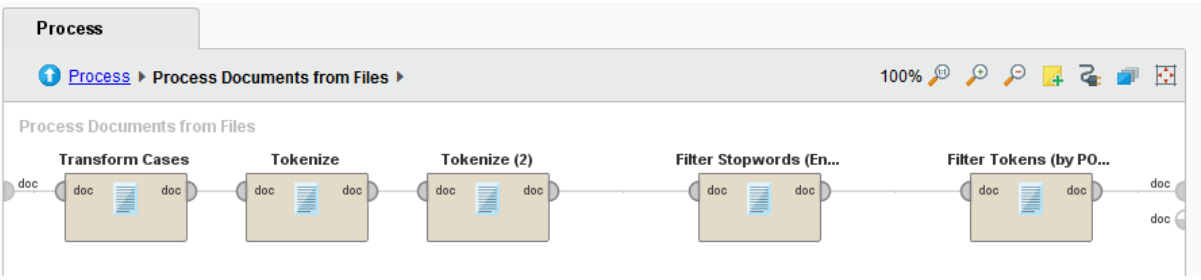

Figure 1. Documents preprocessing, tokenization and vector index computing

According to the TF-IDF,

- ameer (0.49), exemption (0.22), ali (0.16), contraindications (0.15), and ameer's (0.14) were the top 5 terms in the antivax1.txt without pruning (Table 6),
- exemption (0.33), contraindications (0.22), seizure (0.20), — (0.20), and school (0.18) were the top 5 terms in the antivax1.txt after pruning (Table 6),
- pregnancy (0.42), whooping (0.30), cough (0.24), flu (0.19), and women (0.18) were the top 5 terms in the provax1.txt without pruning (Table 6),
- pregnancy (0.48), whooping (0.35), cough (0.28), flu (0.21), and women (0.21) were the top 5 terms in the provax1.txt after pruning (Table 6),
- intradermal (0.68), adjuvanted (0.29), elderly (0.24), srh (0.21), and seasonal (0.14) were the top 5 terms in the neutral1.txt without pruning (Table 7),
- adjuvanted (0.52), elderly (0.43), seasonal (0.25), non-inferiority (0.24), and µg (0.17) were the top 5 terms in the neutral1.txt after pruning (Table 7),
- posh (0.59), alix (0.50), ubiquitination (0.29), hiv-1 (0.21), and release (0.18) were the top 5 terms in the irrelevant1.txt without pruning (Table 7),
- release (0.61), gag (0.25), enhances (0.25), substrate (0.25), function (0.14) were the top 5 terms in the irrelevant1.txt after pruning (Table 7).

The TF-IDF also revealed some punctuation, characters and terms such as *( ) " ' “ ”• ” n't i'm 's* that are noise should be filtered out.

Table 6. Top 20 terms of non-pruned and pruned antivax1.txt and provax1.txt documents based on TF-IDFs

antivax1.txt				**provax1.txt**			
Non-pruned		**Pruned**		**Non-pruned**		**Pruned**	
Term	**TF-IDF**	**Term**	**TF-IDF**	**Term**	**TF-IDF**	**Term**	**TF-IDF**
ameer	0.49	exemption	0.33	pregnancy	0.42	pregnancy	0.48
exemption	0.22	contraindications	0.22	whooping	0.30	whooping	0.35
ali	0.16	seizure	0.20	cough	0.24	cough	0.28
contraindications	0.15	—	0.20	flu	0.19	flu	0.21
ameer's	0.14	school	0.18	women	0.18	women	0.21
contraindication	0.14	seizures	0.17	tdap	0.15	tdap	0.18
seizure	0.14	acip	0.15	non-pregnant	0.15	pregnant	0.17
—	0.13	court	0.15	“whoop”	0.15	coughing	0.15
school	0.12	york	0.14	pregnant	0.14	college	0.15
hamideh	0.12	neurologist	0.13	coughing	0.13	nurse	0.13
seizures	0.12	precautions	0.13	college	0.13	obstetricians	0.12
medical	0.11	denied	0.10	acog	0.12	complications	0.11
acip	0.10	valid	0.10	nurse	0.11	babies	0.11
court	0.10	's	0.10	obstetricians	0.10	reviewed	0.10
york	0.10	case	0.10	complications	0.10	i'm	0.10
gastaut	0.09	advisory	0.10	babies	0.09	suffer	0.10
hamidehs	0.09	doctors	0.10	vaccinated	0.09	highest	0.10
lennox	0.09	department	0.09	reviewed	0.09	hospitalization	0.09
neurologist	0.09	n't	0.09	i'm	0.09	pass	0.09
precautions	0.09	sued	0.09	suffer	0.09	sound	0.09

Table 7. Top 20 terms of non-pruned and pruned neutral1.txt and irrelevant1.txt documents based on TF-IDFs

neutral1.txt				irrelevant1.txt			
Non-pruned		Pruned		Non-pruned		Pruned	Pruned
Term	TF-IDF	Term	TF-IDF	Term	TF-IDF	Term	TF-IDF
intradermal	0.68	adjuvanted	0.52	posh	0.59	release	0.61
adjuvanted	0.29	elderly	0.43	alix	0.50	gag	0.25
elderly	0.24	seasonal	0.25	ubiquitination	0.29	enhances	0.25
srh	0.21	non-inferiority	0.24	hiv-1	0.21	substrate	0.25
seasonal	0.14	µg	0.17	release	0.18	function	0.14
non-inferiority	0.13	titres	0.16	mutant	0.15	protein	0.14
µg	0.10	randomised	0.15	hiv-1δptap	0.10	daniel	0.13
titres	0.09	n	0.14	l-domain	0.10	exploring	0.13
randomised	0.08	strains	0.13	gag	0.07	inactive	0.13
influenza	0.08	immunogenicity	0.11	enhances	0.07	interacts	0.13
n	0.08	method	0.11	substrate	0.07	x	0.13
strains	0.07	dose	0.10	(ypxnl-dependent	0.05	derived	0.12
vaccine	0.07	methods	0.10	://bmcbiochem	0.05	facilitate	0.12
adjust	0.07	demonstrated	0.10	adaptor	0.05	mediated	0.12
arnou	0.07	antibody	0.10	alix-augmentation	0.05	membrane	0.12
baseline	0.07	single	0.10	augmentation	0.05	substantially	0.12
comparable	0.07	ci	0.09	augments	0.05	wild	0.12
comparative	0.07	haemagglutination	0.09	auxiliary	0.05	central	0.11
covariance	0.07	haemagglutinin	0.09	bannert	0.05	functional	0.11
damme	0.07	non-inferior	0.09	biogenesis	0.05	modified	0.11

The highest TF-IDF was extracted for each document using the MAX function of Excel. The maximum TF-IDF plotted on Figure 2 per document whether non-pruned or pruned, shows a high variability. Although maximum TF-IDFs follow the same pattern for non-pruned as well as pruned documents, the pruning yields higher TF-IDFs as compared to non-pruning.

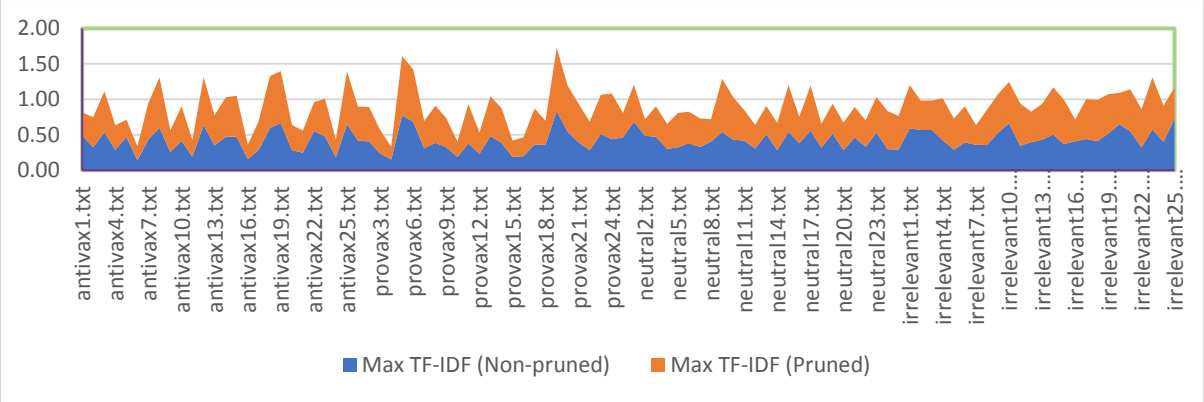

Figure 2. Maximum TF-IDF per document (pruned and non-pruned)

3.4. Text visualization

Text visualization was performed on the training set of each class of document to identify words that appear to be the most popular.

After reading each class of documents in the Process Documents from Files operator (TF-IDF vector creation, pruning below 3% and above 30%, sub processes as on Figure 1) and transforming word list to data, word clouds were created in R Scripting operator (Figure 3) using the following code:

```
library("wordcloud2")
library("wordcloud")
library("RColorBrewer")
png("C:/Users/path/antivax_wordcloud.png", width=800, height=800)
wordcloud(words = data$word, freq = data$total, min.freq = 2, max.words=500,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
dev.off()
return (data)
```

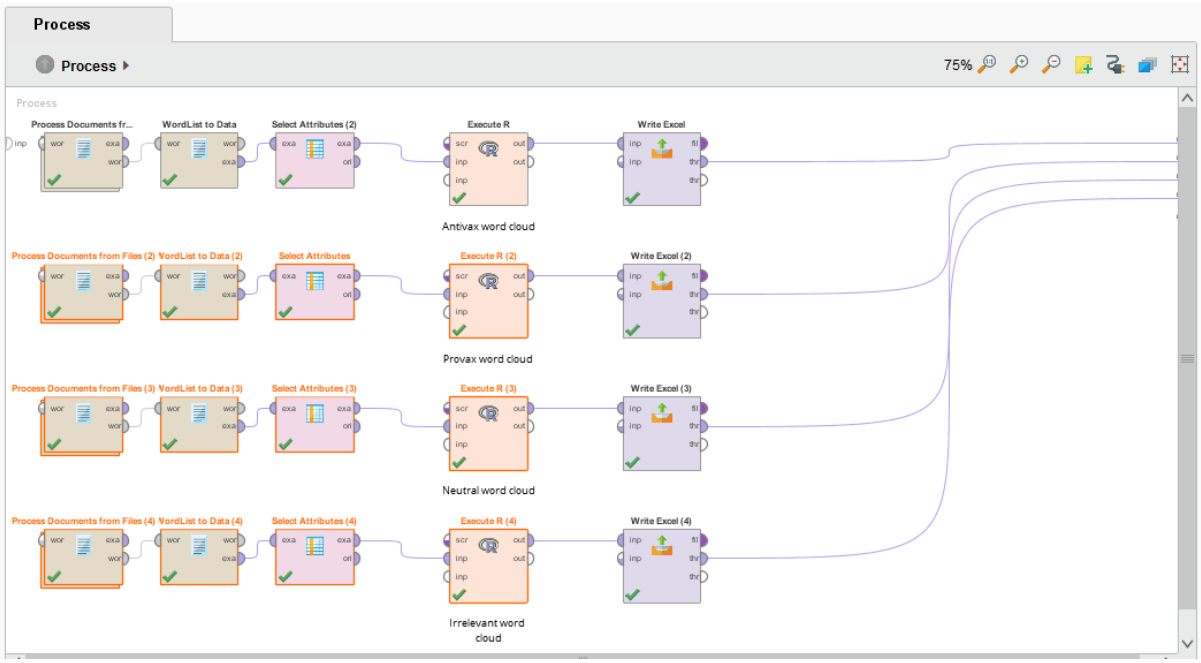

Figure 3. RapidMiner processes to produce word clouds from documents in each of the four categories

Word cloud of documents in the training set documents shows that:

- the terms *bills, cancer, hb, passed, and fetal,* are the top-5 most popular words in the Antivax category (Figure 4),
- the terms *shingles, and influenza, autism, hepatitis,* are the top-5 most popular words in the Provax category (Figure 5),
- the terms *pandemic, tiv, hiv, qiv, and groups,* are the top-5 most popular words in the Neutral category (Figure 6),
- the terms *domain, saps, expression, kinase, and phosphorylation* are the top-5 most popular words in the Irrelevant category (Figure 7).

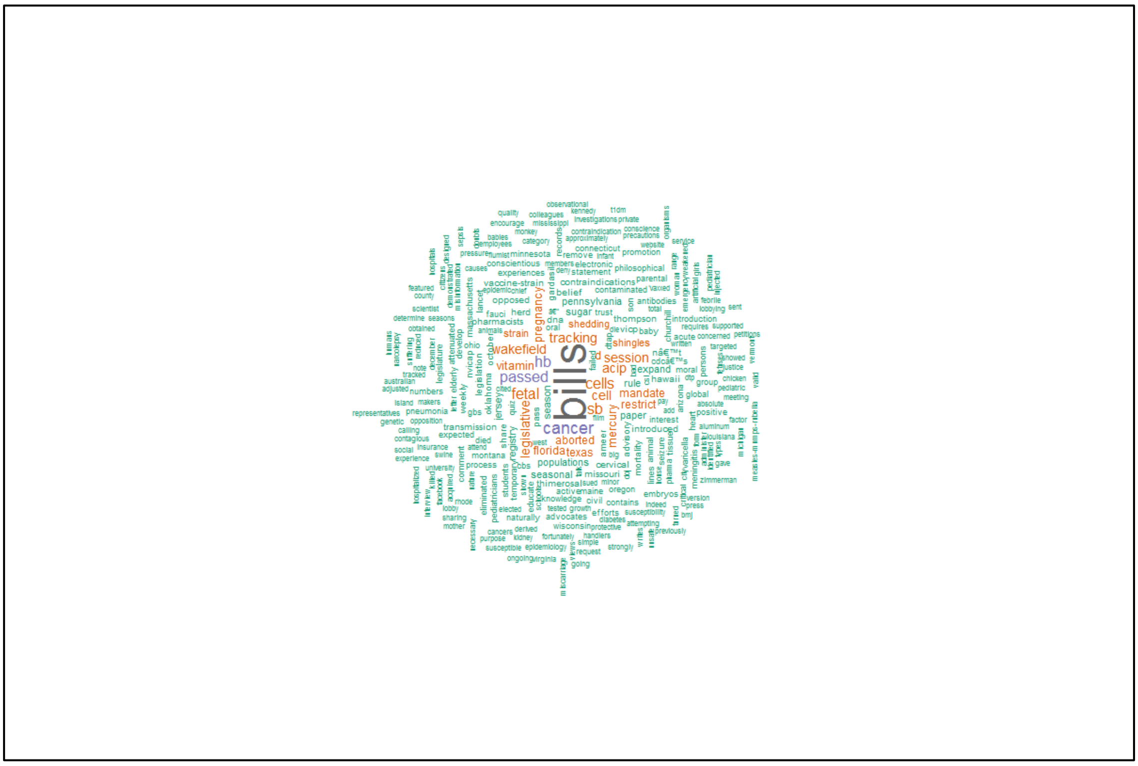

Figure 4. Word cloud of documents in the Antivax training set documents

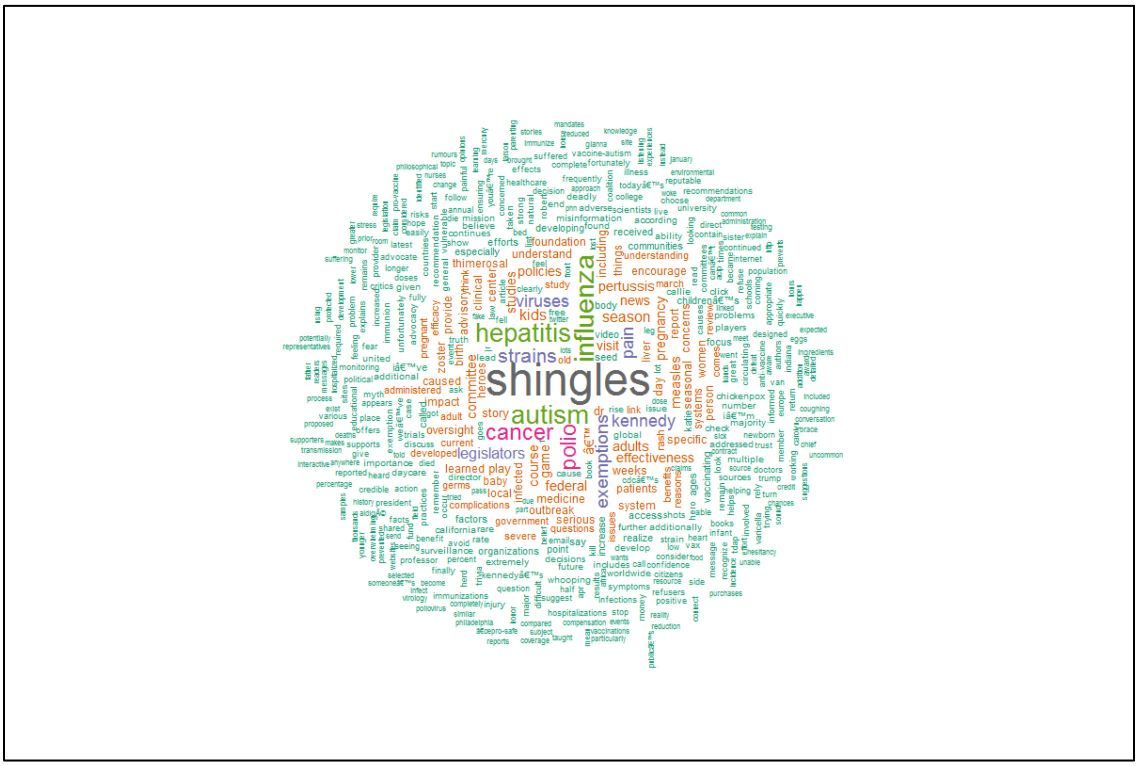

Figure 5. Word cloud of documents in the Provax training set documents

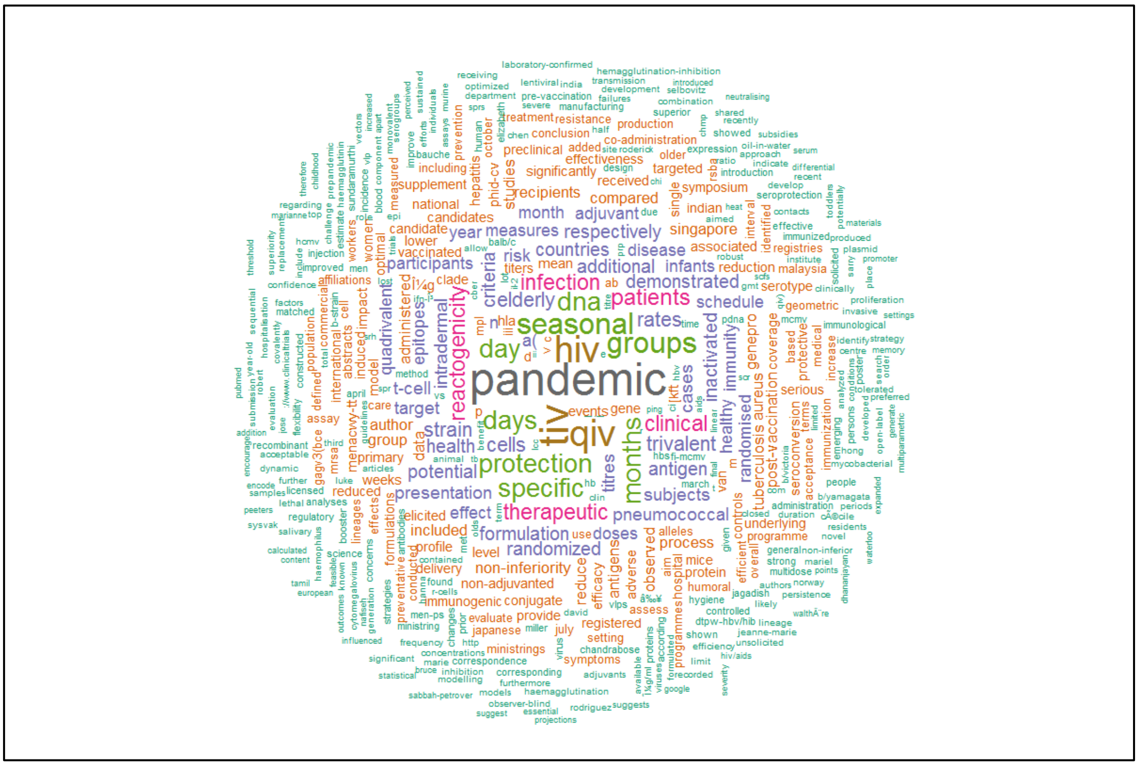

Figure 6. Word cloud of documents in the Neutral training set documents

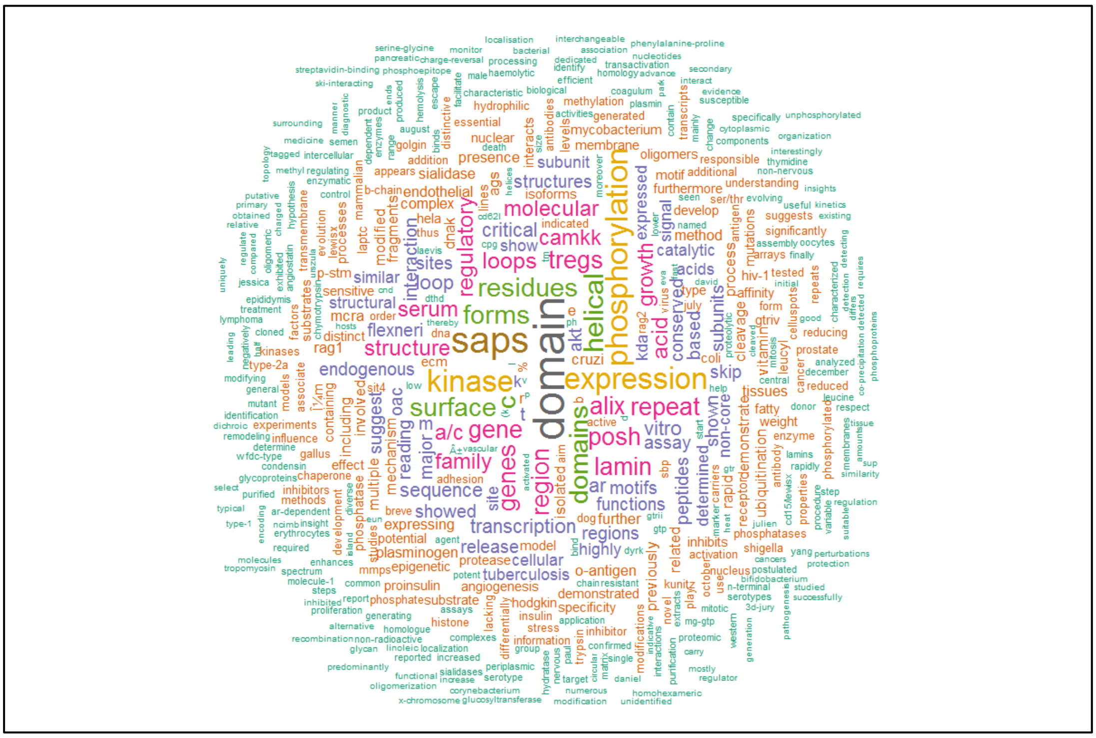

Figure 7. Word cloud of documents in the Irrelevant training set documents

IV. Data preparation

The data preparation will use the operators presented on Figure 1 (limited to Transform Cases, Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English)), and new added operators such as Replace Token, Filter Stopwords (Dictionary), and Stem, with several combinations and over several iterations to identify the combination that yields the best performance (Average withing distance for clustering, and Accuracy, Precision, and Recall for classification).

V. Modeling

5.1. Clustering

K-mean was chosen for the documents clustering. Following the document processing operator, the clustering, performance, and log operators were added inside a parameters optimization operator to determine the optimal number of clusters (K value) along with the best measurement type and main performance criterion (Figure 8).

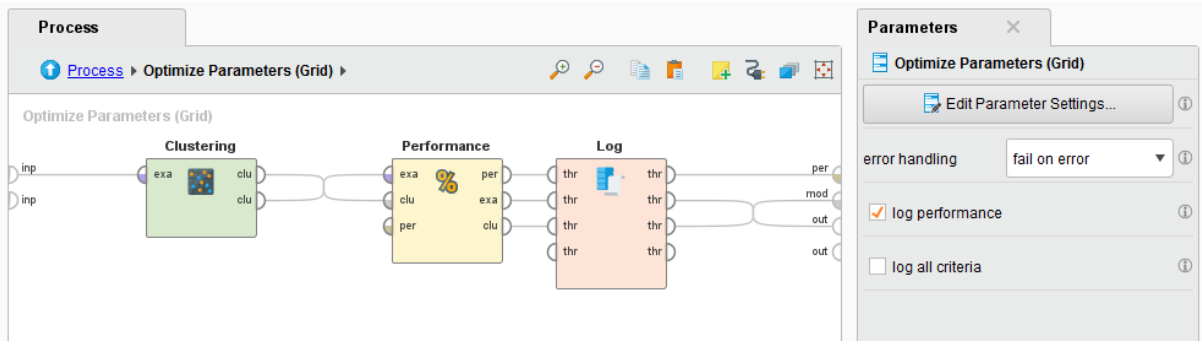

Figure 8. RapidMiner process tom determine optimal cluster parameters

The parameter tuning shows that 4 is the optimal number of clusters and the Davies Bouldin index is the best performance measure.

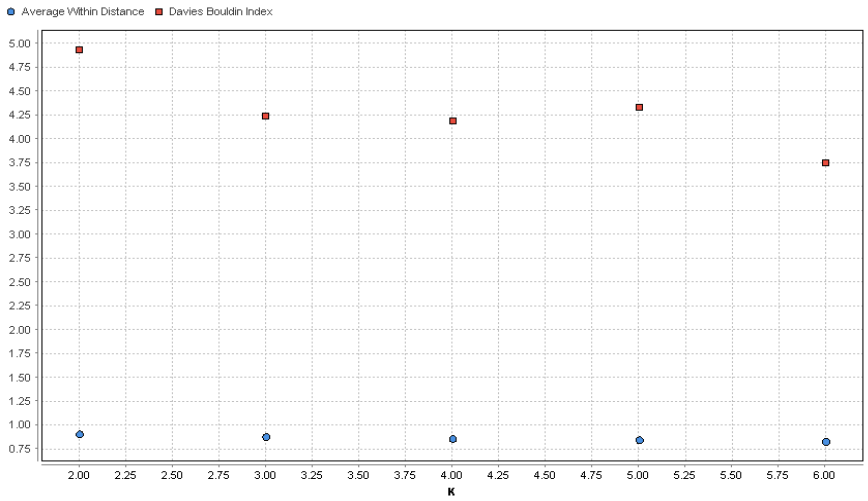

Figure 9. Elbow plot to determine the optimal value of K

Executing the clustering operator with K=4 generates a model as follows:

Cluster 0: 25 items

Cluster 1: 20 items

Cluster 2: 25 items

Cluster 3: 30 items

Total number of items: 100

This shows that the clusters are approximately of the same size, although the 4th cluster is of a little bigger size and the 2nd cluster is of a little smaller size. The performance measure (Average within centroid distance) also shows that the clusters centroids are at approximately the same distance from each other:

Avg. within centroid distance: -0.863

Avg. within centroid distance_cluster_0: -0.823

Avg. within centroid distance_cluster_1: -0.831

Avg. within centroid distance_cluster_2: -0.887

Avg. within centroid distance_cluster_3: -0.896

Davies Bouldin: -4.414

Adding singular value decomposition (SVD) with 2 dimensions to the process, 3D visualization shows the 3 clusters. It can be observed that there is some overlap between cluster 2 and the others.

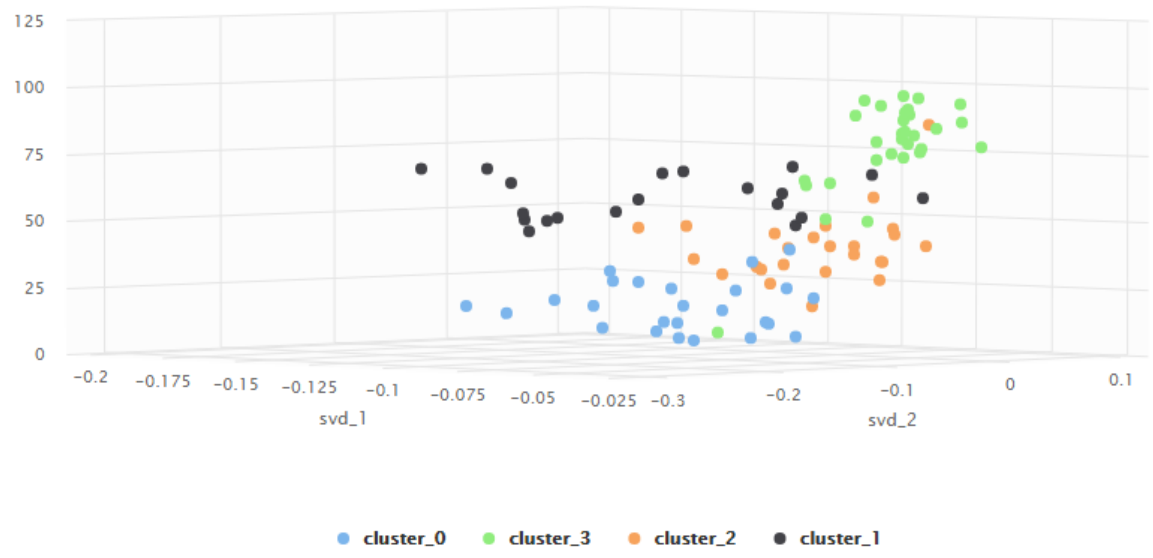

Figure 10. 3D visualization of clusters with SVD

Two classification algorithms (K-NN) and Decision tree (DT) with default parameters were applied to the clustering algorithm developed (Figure 11). The DT generated is presented in Figure 12.

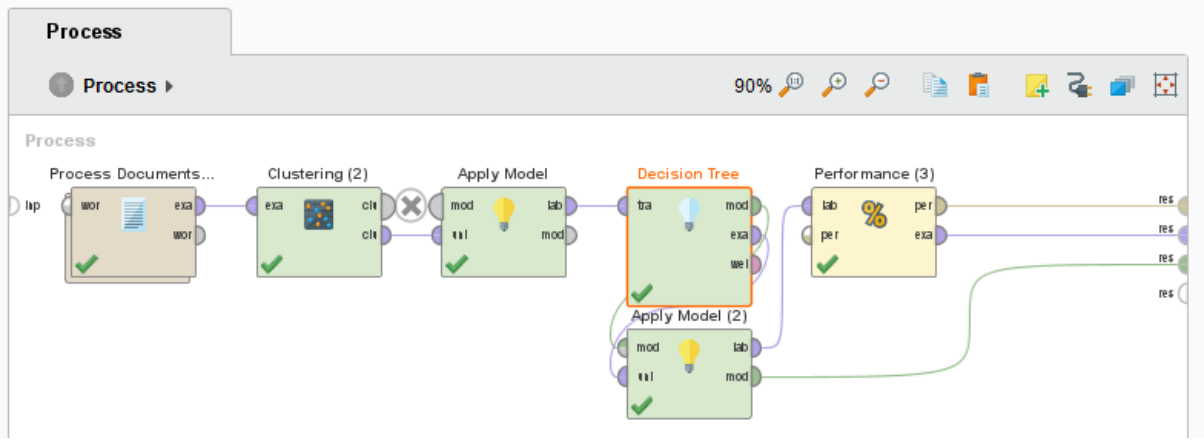

Figure 11. RappidMiner process to apply classification algorithms to the clustering

Table 8. Performance of K-NN and Decision tree algorithms applied to the clustering

		Precision (%)	Recall (%)	Accuracy (%)
K-NN				
	Cluster 0	89.29	100.00	93.00
	Cluster 1	93.75	100.00	
	Cluster 2	95.24	80.00	
	Cluster 3	94.74	90.00	
Decision tree				
	Cluster 0	100.00	96.00	74.00
	Cluster 1	53.57	100.00	
	Cluster 2	0.00	0.00	
	Cluster 3	100.00	100.00	

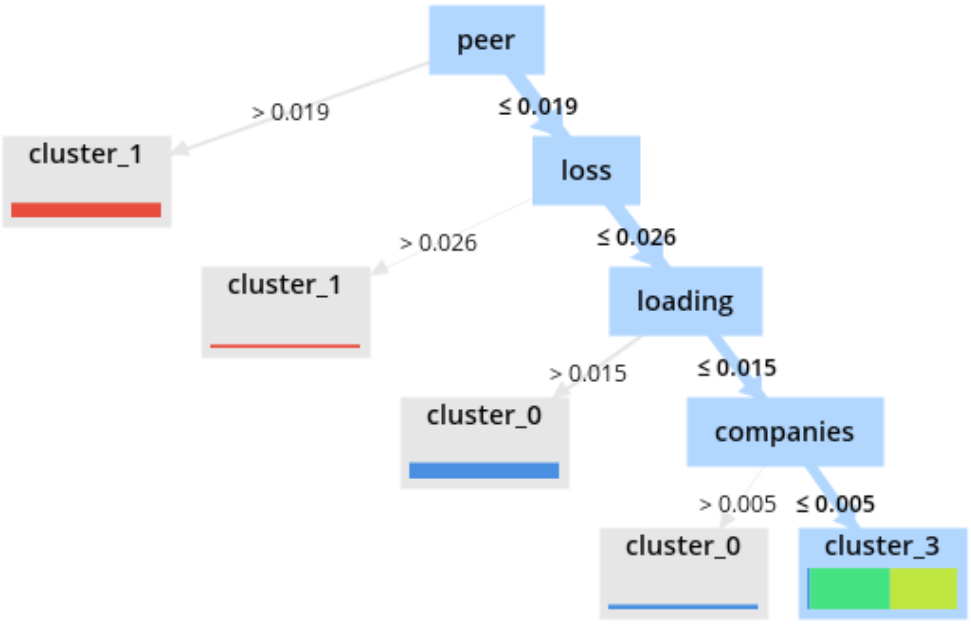

Figure 12. Decision tree generated from clustering

The evaluation the clustering on the two classification algorithms shows that the K-NN has a better performance with 93.00% accuracy, recalls between 80%-100% and precisions between 89.29%-95.24%.

5.2. Classification

Five classification algorithms were applied on the labeled dataset: K-NN, Naïve Bayes, SVM, Decision Tree (DT), and Deep Learning (DL). The models were developed using the Process Document from File operator, connected to the Cross-Validation operator to perform 10-fold cross validation (Figure 13). Process Document from File operator contained text processing operator presented on Figure 1.

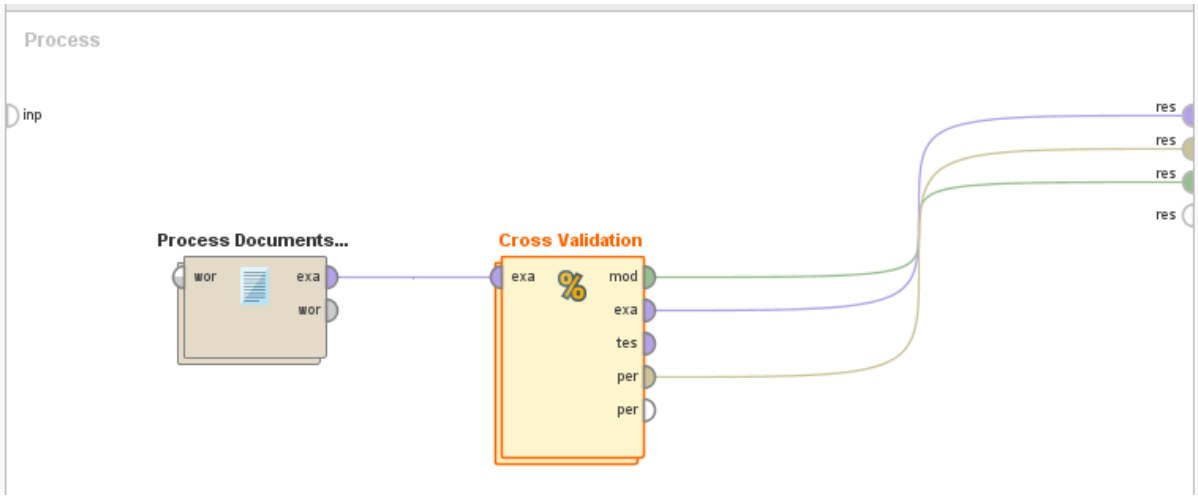

Figure 13. Overview of the classification process

The Cross Validation operator contained the classification operator on the left (K-NN, Naïve Bayes, or SVM), and Apply Model and Performance operator on the right (Figure 14).

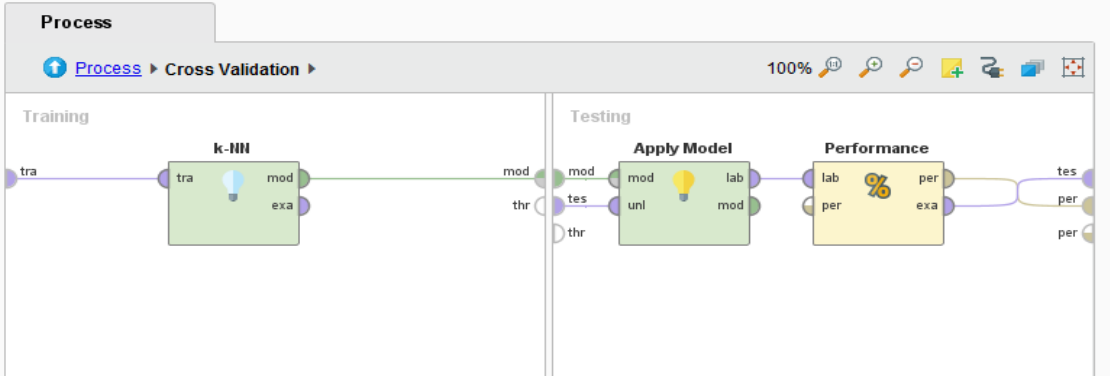

Figure 14. Overview of operators in the Cross-Validation operator to perform classification

With the setup presented in Figure 8 and Figure 9, the development of various classification algorithms only depends on the classifier and the parameters tuning. Thus, the next sections will be dedicated to finding the best parameters to yield the highest accuracies for K-NN, Naïve Bayes, and SVM.

5.2.1. K-Nearest Neighbors (K-NN)

The first step taken with K-NN classification is finding the best schema for creating the word vectors and optimal k value.

After putting the process presented on Figure 8 in a Optimize Parameters operator, TF-IDF vectors creation schema and k=11 were found to be one of the best combinations to produce an optimal accuracy (Accuracy = 92%) (Figure 15).

iteration	Process Document...	k-NN.k	accuracy ↓
25	TF-IDF	11	0.920
17	TF-IDF	9	0.920
29	TF-IDF	12	0.920
37	TF-IDF	14	0.900
9	TF-IDF	7	0.900
21	TF-IDF	10	0.900
3	Term Occurrences	5	0.900
7	Term Occurrences	6	0.900
33	TF-IDF	13	0.900
5	TF-IDF	6	0.890
41	TF-IDF	15	0.890
1	TF-IDF	5	0.880
13	TF-IDF	8	0.880
11	Term Occurrences	7	0.870
15	Term Occurrences	8	0.870

Figure 15. Results of the vector creation schema and k parameters tuning

From now, the two parameters are kept constant (TF-IDF, k = 11) and perceptual pruning values are varied to find the optimal accuracy. The pruning step showed that ignoring words that appear in less than 4% and in more than 35% of all documents yield an optimal accuracy (94%) (Figure 16).

iteration	Process Docume...	Process Doc...	acc... ↓
39	9	34	0.940
42	4	35	0.940
84	6	40	0.940
48	10	35	0.930
49	3	36	0.930
32	10	33	0.930
56	10	36	0.930
15	9	31	0.930
19	5	32	0.930
21	7	32	0.930
70	8	38	0.930
75	5	39	0.930
87	9	40	0.930
2	4	30	0.920
47	9	35	0.920

Figure 16. Results of the K-NN with various combinations of pruning values

The combination of the four parameters identified above (TF-IDF, k = 11, prune below 4% and above 35%, and Transform Cases, Tokenize (linguistic sentences and linguistic tokens), and Filter Stopwords (English) as preprocessing steps) was considered as a baseline iteration, and produced an accuracy of 94% with 83%, 95%, 100% precision, and 100%, 80%, 96%, and 100% recall for antivax, provax, neutral and irrelevant classes respectively (Figure 17).

accuracy: 94.00% +/- 8.43% (micro average: 94.00%)

	true antivax	true provax	true neutral	true irrelevant	class precision
pred. antivax	25	5	0	0	83.33%
pred. provax	0	20	1	0	95.24%
pred. neutral	0	0	24	0	100.00%
pred. irrelevant	0	0	0	25	100.00%
class recall	100.00%	80.00%	96.00%	100.00%	

Figure 17. Performance of baseline iteration of the k-NN algorithm with TF-IDF, k = 11, and prune below 4% and above 35%, and Transform Cases, Tokenize (linguistic sentences and linguistic tokens), and Filter Stopwords (English) as preprocessing steps

In the next steps, various text processing operator were added inside the Process Document from File operator presented on Figure 8, to assess if there will be any improvement to the k-NN performance metrics (Accuracy, Precision, and Recall). The summary of the iterations is as follows:

- 1st Iteration Baseline*
- 2nd Iteration Baseline, Replace Tokens
- 3rd Iteration Baseline, Replace Tokens, Filter Tokens (by POS Tags**)
- 4th Iteration Baseline, Replace Tokens, Porter Steamer
- 5th Iteration Baseline, Replace Tokens, Lovins Steamer
- 6th Iteration Baseline, Replace Tokens, Snowball Steamer
- 7th Iteration Baseline, Replace Tokens, Snowball Steamer, Filter Tokens (by POS Tags)
- 8th Iteration Baseline, Replace Tokens, Generate n-Grams*** (Terms)
- 9th Iteration Baseline, Replace Tokens, Snowball Steamer, Generate n-Grams (Terms)
- 10th Iteration Baseline, Snowball Steamer
- 11th Iteration Baseline, Generate n-Grams (Terms)

(*) Vector creation parameters: TF-IDF, prune below 4% and above 35%); k = 11; preprocessing operators: Transform Cases, Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English).

(**) Expression: JJ.*|N.*|R.*|V.*

(***) Maximal length: 2.

The evaluation of the 11 iterations based on Accuracy, Class precision and Class recall is presented in Table 9. These results show that the best performance is yield by the 11th iteration: TF-IDF, k = 11, and prune below 4% and above 35%, and Transform Cases, Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English), and Generate n-Grams (Terms). The model was accurate at 97%, with a precision of 89.29% for the Antivax class, and 100% for the Provax, Neutral, and Irrelevant classes, and 88% of recall for the Provax class, and 100% for Antivax, Neutral, and Irrelevant classes (Table 9).

Table 9. Results of the evaluation of the 11 k-NN iterations based on Accuracy, Precision and Recall

Iteration #	Accuracy	Precision				Recall			
		Antivax	Provax	Neutral	Irrelevant	Antivax	Provax	Neutral	Irrelevant
1st Iteration	94.00	83.33	95.24	100.00	100.00	100.00	80.00	96.00	100.00
2nd Iteration	91.00	75.68	94.44	100.00	100.00	100.00	68.00	96.00	100.00
3rd Iteration	90.00	78.12	94.74	10.00	92.59	100.00	72.00	88.00	100.00
4th Iteration	90.00	75.00	94.44	100.00	96.15	96.00	68.00	96.00	100.00
5th Iteration	93.00	82.14	90.91	100.00	100.00	92.00	80.00	100.00	100.00
6th Iteration	89.00	74.19	89.47	100.00	96.15	92.00	68.00	96.00	100.00
7th Iteration	88.00	74.19	85.00	100.00	96.15	92.00	68.00	92.00	100.00
8th Iteration	93.00	80.65	100.00	100.00	92.59	100.00	80.00	92.00	100.00

9th Iteration	93.00	83.33	100.00	100.00	92.59	100.00	80.00	92.00	100.00
10th Iteration	93.00	83.33	100.00	100.00	92.59	100.00	80.00	92.00	100.00
11th Iteration	97.00	89.29	100.00	100.00	100.00	100.00	88.00	100.00	100.00

5.2.2. Naïve Bayes (NB)

The same process was used as presented on Figure 18 and Figure 19, but the K-NN operator was replaced by the Naïve Bayes operator. The process was placed in a parameter optimizer to determine the best ones. As shown on Figure 13, the highest accuracy (79%) is yield by the combination of Binary Term Occurrences and Laplace correction.

iteration	Process Documents from ...	Naive Bayes.laplace_correction	acc... ↓
4	Binary Term Occurrences	true	0.790
8	Binary Term Occurrences	false	0.760
2	Term Frequency	true	0.660
1	TF-IDF	true	0.650
6	Term Frequency	false	0.650
7	Term Occurrences	false	0.650
5	TF-IDF	false	0.630
3	Term Occurrences	true	0.610

Figure 18. Results of the Naive Bayes algorithm with vector creation schemas and Laplace correction

The optimal pruning result is yielded by ignoring words that appear in less than 6% and in more than 46% of all documents (Figure 19).

iteration	Process Docum...	Process Documen...	accuracy ↓
132	6	46	0.870
29	7	33	0.870
35	5	34	0.870
74	4	39	0.870
11	5	31	0.860
100	6	42	0.860
147	5	48	0.860
23	9	32	0.860
151	9	48	0.860
153	3	49	0.860
167	9	50	0.860
125	7	45	0.860
126	8	45	0.860
49	3	36	0.860
130	4	46	0.860

Figure 19. Results of the NB with various combinations of pruning values

The combination of the parameters determined above, which are kept constant (Binary Term Occurrences, and prune below 6% and above 46%) and Transform Cases, Tokenize (linguistic sentences and linguistic tokens), and Filter Stopwords (English) is used as baseline. The same iterations as listed in Section 5.1.1 are performed to determine the optimal preprocessing approach.

Table 10. Results of the evaluation of the 11 Naïve Bayes iterations based on Accuracy, Precision and Recall

Iteration	Accuracy	Precision				Recall			
		Antivax	Provax	Neutral	Irrelevant	Antivax	Provax	Neutral	Irrelevant
1st Iteration*	89.00	69.44	100.00	100.00	100.00	100.00	80.00	76.00	100.00
2nd Iteration	91.00	73.53	100.00	100.00	100.00	100.00	84.00	80.00	100.00
3rd Iteration	88.00	67.57	100.00	100.00	100.00	100.00	76.00	76.00	100.00
4th Iteration	88.00	73.53	86.36	100.00	100.00	100.00	76.00	76.00	100.00
5th Iteration	85.00	64.10	94.44	100.00	100.00	100.00	68.00	76.00	96.00
6th Iteration	59.00	38.46	83.33	100.00	100.00	100.00	20.00	52.00	64.00
7th Iteration	87.00	73.53	80.95	100.00	100.00	100.00	68.00	80.00	100.00
8th Iteration	95.00	86.21	95.83	100.00	100.00	100.00	92.00	88.00	100.00
9th Iteration	96.00	92.59	92.31	100.00	100.00	100.00	96.00	88.00	100.00
10th Iteration	91.00	75.76	95.24	100.00	100.00	100.00	80.00	84.00	100.00

11th Iteration	98.00	92.59	100.00	100.00	100.00	100.00	96.00	96.00	100.00

(*) Baseline. Vector creation parameters: Binary Term Occurrences, and prune below 6% and above 46%; preprocessing operators: Transform Cases, Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English).

The evaluation of the 11 iterations based on Accuracy, Class precision and Class recall is presented in Table 10. These results show that the optimal performance is yielded by the 11th iteration: Binary Term Occurrences, and prune below 6% and above 46%, and Transform Cases, Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English), and Generate n-Grams (Terms). The model was accurate at 98%, with a precision of 92.59% for the Antivax class, and 100% for the Provax, Neutral, and Irrelevant classes, and 88% of recall for the Provax class, and 100% for Antivax, Neutral, and Irrelevant classes (Table 8).

5.2.3. Decision tree (DT)

The same process presented on Figure 8 and Figure 9 was used, but the K-NN operator was replaced by the Decision Tree operator. The process was placed in an optimizer to determine the best parameter. As shown on Figure 13, the highest accuracy (99%) is yielded by the combination of Term Frequency and Gini Index (Figure 20).

iteration	Process Documents ...	Decision Tree.c...	acc... ↓
10	Term Frequency	gini_index	0.990
1	TF-IDF	gain_ratio	0.980
12	Binary Term Occurren...	gini_index	0.980
4	Binary Term Occurren...	gain_ratio	0.980
7	Term Occurrences	information_gain	0.960
8	Binary Term Occurren...	information_gain	0.960
6	Term Frequency	information_gain	0.960
2	Term Frequency	gain_ratio	0.960
11	Term Occurrences	gini_index	0.960
16	Binary Term Occurren...	accuracy	0.950
9	TF-IDF	gini_index	0.950
5	TF-IDF	information_gain	0.950
13	TF-IDF	accuracy	0.950
15	Term Occurrences	accuracy	0.950
3	Term Occurrences	gain_ratio	0.940

Figure 20. Results of the DT algorithm with vector creation schemas and criterion

The tuning of maximal depth, confidence, and minimal gain shows that several combinations of these 3 parameters can yield an optimal accuracy. Maximal depth = 10, confidence = 0.1, and minimal gain = 0.03 were arbitrary chosen for the next steps (Figure 21).

iteration	Decision Tree.maximal_depth	Decision Tree.confidence	Decision Tree.mini...	acc... ↓
80	11	0.400	0.030	1
45	12	0.300	0.020	1
88	13	0.500	0.030	1
61	10	0.100	0.030	1
25	10	0.500	0.010	1
104	11	0.300	0.040	1
147	12	0.500	0.050	1
76	13	0.300	0.030	0.990
77	14	0.300	0.030	0.990
78	15	0.300	0.030	0.990
79	10	0.400	0.030	0.990
5	14	0.100	0.010	0.990
42	15	0.200	0.020	0.990
81	12	0.400	0.030	0.990
50	11	0.400	0.020	0.990

Figure 21. Results of the DT algorithm depending on maximal depth, confidence, and minimal gain

While keeping Maximal depth = 10, confidence = 0.1, and minimal gain = 0.03 constant and pruning vector creation step, one of the top performances was obtained by ignoring words that appear in less than 10% and in more than 40% of all documents (Figure 22).

iteration	Proces...	Process Doc...	acc... ↓
45	7	35	1
80	10	39	1
88	10	40	1
24	10	32	1
61	7	37	1
114	4	44	1
104	10	42	1
1	3	30	0.990
2	4	30	0.990
67	5	38	0.990
35	5	34	0.990
36	6	34	0.990
37	7	34	0.990
6	8	30	0.990

Figure 22. Results of the DT with various combinations of pruning values

Based on the steps described above, the optimal parameters and processes were determined and can be summarized as follows:

- Text processing: Transform Cases, Tokenize (linguistic sentences and linguistic tokens), and Filter Stopwords (English),
- Vector creation: Term Frequency and pruning less than 10% and more than 40%,
- DT: Gini Index criterion, Maximal depth = 10, confidence = 0.1, and minimal gain = 0.03.

With these parameters and processes, the DT developed (Figure 23) yielded an accuracy of 99% with 100% of precision for antivax, provax and irrelevant classes, and 96.15% for neutral class; 100% recall for provax, neutral and irrelevant classes, and 96% for antivax class (Figure 24).

Given these results, decision was taken to not perform any other preprocessing iteration.

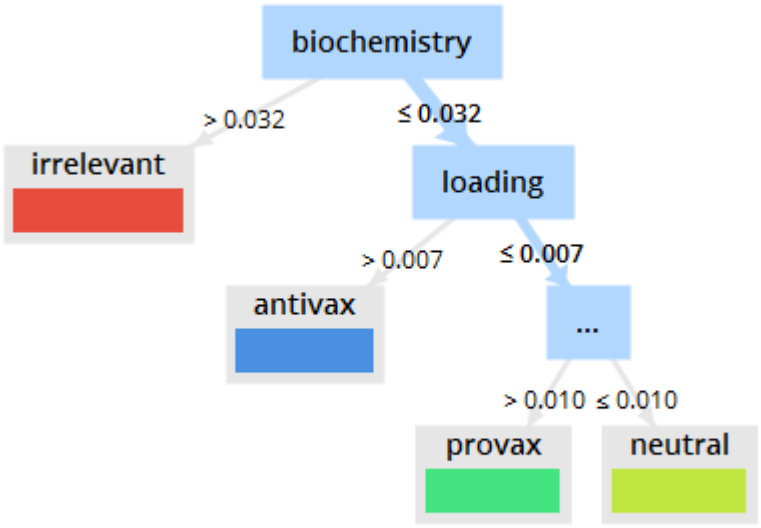

Figure 23. DT developed with the optimal parameters determined

accuracy: 99.00% +/- 3.16% (micro average: 99.00%)

	true antivax	true provax	true neutral	true irrelevant	class precision
pred. antivax	24	0	0	0	100.00%
pred. provax	0	25	0	0	100.00%
pred. neutral	1	0	25	0	96.15%
pred. irrelevant	0	0	0	25	100.00%
class recall	96.00%	100.00%	100.00%	100.00%	

Figure 24. Performance of the DT with the optimal parameters determined

5.2.4. Support vector machine (SVM)

The same process presented on Figure 8 and Figure 9 was used, but the K-NN operator was replaced by the SVM operator. Given the multiclass nature of the dataset, the Library for Support Vector Machines (LibSVM) was specifically used. The process was placed in an optimizer to determine the best parameters. Term Frequency vector creation and nu-SVC SVM type were determined as the optimal parameters (Figure 25).

iteration	SVM (2).svm_ty...	Process Documents from...	acc... ↓
4	nu-SVC	Term Frequency	1
2	nu-SVC	TF-IDF	0.960
7	C-SVC	Binary Term Occurrences	0.880
8	nu-SVC	Binary Term Occurrences	0.870
1	C-SVC	TF-IDF	0.400
3	C-SVC	Term Frequency	0.370
5	C-SVC	Term Occurrences	0.280
6	nu-SVC	Term Occurrences	0.270

Figure 25. Results of the SVM with various SVM types and document vector creation schemas

With these parameters and processes, the SVM developed yielded 100% accuracy with 100% precision and recall for all four classes (Figure 26).

Given these results, decision was taken to not proceed with further iterations.

accuracy: 100.00% +/- 0.00% (micro average: 100.00%)

	true antivax	true provax	true neutral	true irrelevant	class precision
pred. antivax	25	0	0	0	100.00%
pred. provax	0	25	0	0	100.00%
pred. neutral	0	0	25	0	100.00%
pred. irrelevant	0	0	0	25	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	

Figure 26. Performance of the SVM with the optimal parameters determined

5.2.5. Neural network –deep learning (DL)

The k-NN operator was replaced by de deep learning (DL) operator in the process presented on Figure 8 and Figure 9. The DL default parameters were kept and word vector creation parameters and text processing steps were varied to develop the optimal algorithm.

Term Occurrences was found to the optimal vector creation schema (Figure 27).

iteration	Process Documents from Fil...	acc... ↓
3	Term Occurrences	0.950
4	Binary Term Occurrences	0.890
1	TF-IDF	0.780
2	Term Frequency	0.670

Figure 27. Results of the DL algorithm with various vector creation schemas

The tuning of the pruning at vector creation steps showed that pruning below 4% and above 30% could improve the accuracy from 95% to 97%. Thus, Term Occurrences, pruning below 4% and above 30%, with baseline text processing, yield an accuracy of 97% with 96% precision for the Antivax class, 100% for the Provax and Neutral class, and 92% for the Irrelevant class; 100% recall for the Antivax and Irrelevant class, 92% for the Provax class and 96% for the Neutral class (Figure 28).

accuracy: 97.00% +/- 4.83% (micro average: 97.00%)

	true antivax	true provax	true neutral	true irrelevant	class precision
pred. antivax	25	1	0	0	96.15%
pred. provax	0	23	0	0	100.00%
pred. neutral	0	0	24	0	100.00%
pred. irrelevant	0	1	1	25	92.59%
class recall	100.00%	92.00%	96.00%	100.00%	

Figure 28. Performance of the DL with prune below 4 and above 30

5.2.6. Summary of the optimal models

The summary of optimal models developed above is presented in Table 10. This summary shows that all models have high accuracies, precisions and recalls, but only SVM has 100% accuracy, precision and recall for all document classes. Although SVM has the optimal performance, this project does not evaluate if it is related to overfitting. Also, the processing time of SVM (1:11 min) longer (1.5 time) as compared to the processing time of DT (0:48 min) which has 99% accuracy. Based on this summary, the choice of a model that would be accurate as well as fast would be DT.

Table 11. Summary of optimal models

Algorithm	Parameters	Vector creation	Text processing operators*	Accuracy	Precision *(Antivax, Provax, Neutral, Irrelevant)*	Recall *(Antivax, Provax, Neutral, Irrelevant)*	Execution time (min)
K-NN	k = 11	TF-IDF, prune (4-35)	Generate n-Grams	97%	89%, 100%, 100%, 100%	100%, 88%, 100%, 100%	1:39
NB	Laplace correction	BTO, prune (6 - 46)	Generate n-Grams	98%	93%, 100%, 100%, 100%	100%, 96%, 96%, 100%	1:41
DT	Gini Index, Maximal depth = 10, confidence = 0.1, Minimal gain = 0.03.	TF, prune (10-40)	Baseline	99%	100%, 100%, 100% 96.15%	96% 100%, 100%, 100%	0:48
SVM	nu-SVC	TF	Baseline	100%	100%, 100%, 100%, 100%	100%, 100%, 100%, 100%	1:11
DL	Default	TO, pruning (4 - 30)	Baseline	97%	96%, 100%, 100%, 92%	100%, 92%, 96%, 100%	1:44

(*) Operators added to the baseline (Transform Cases, Tokenize (linguistic sentences and linguistic tokens), Filter Stopwords (English))

5.2.7. Application on unseen data

To apply the models developed on unseen documents, the wordlist created in the document preprocessing operator have been saved in object storing operators. The models have also been saved in the same way (Figure 29).

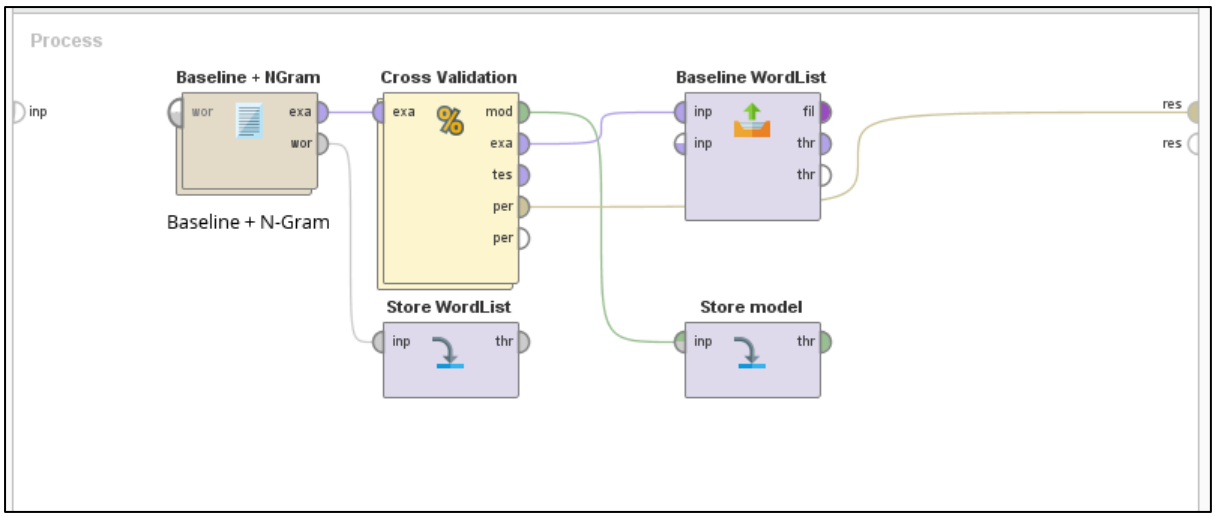

Figure 29. Storage of word list and model

The word lists are then applied to document processing operators containing unseen documents and the saved models applied the unseen processed documents in the model application operator (Figure 30).

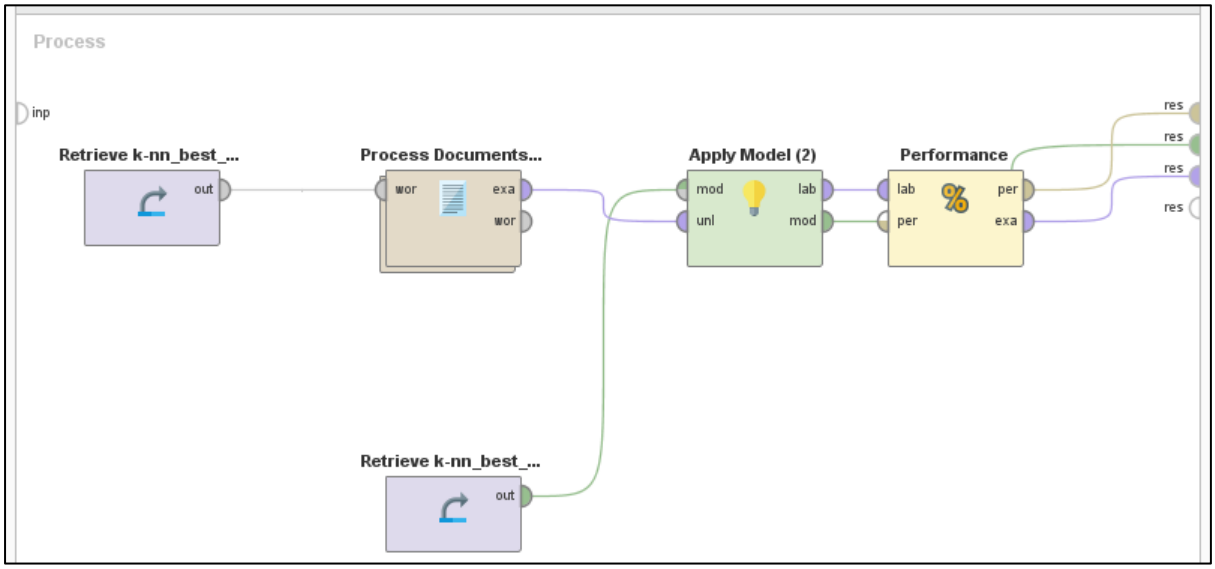

Figure 30. Application of word list and saved models on unseen documents

The summary of the performance of the models on unseen documents is presented in Table 12. This summary shows that the optimal performance is yielded by the NB with 83% accuracy.

Table 12. Performance of models applied on unseen data

Algorithm	Accuracy	Precision *(Antivax, Provax, Neutral, Irrelevant)*	Recall *(Antivax, Provax, Neutral, Irrelevant)*
K-NN	78%	88%, 100%, 70%, 62%	100%, 80%, 50%, 87%
NB	83%	88%, 93%, 70%, 90%	93%, 87%, 93%, 60%
DT	25%	0%, 0%, 25%, 0%	0%, 0%, 100%, 0%
SVM	27%	0%, 0%, 13%, 93%	0%, 0%, 33%, 26%
DL	50%	75%, 39%, 48%, 55%	20%, 33%, 73%, 73%

VI. Conclusion

This report has presented the development of clustering and classification algorithms on a set of documents to detect antivaccine webpages as compared to provaccine, neutral, and irrelevant ones. It demonstrated the possibility to cluster the documents into 3, 4, or 4 groups and apply several classification algorithms on them.

A set of tools and software such as Textise (CPC LLC, Austin, TX), Link Grabber, RapidMiner, Microsoft Excel (Microsoft Corporation, Redmond, WA), and R (The R Project for Statistical Computing, R Core Team) were used. Few preprocessing operators were used as baseline preprocessing parameters to which several other processors were added across multiple iterations to determine the optimal preprocessing workflow. 10-fold cross validation was used for models' evaluation with accuracy, precision and recall as evaluation metrics. All models have high accuracies, precisions and recalls, but only SVM has 100% accuracy, precision and recall for all document classes. However, when the models are applied on unseen documents, the evaluation show that the highest performing is NB with 83% accuracy and was able to correctly predict Antivax, Provax, Neutral, and Irrelevant documents respectively at 88%, 93%, 70%, and 90% among the total predicted one (precision), and correctly predict at 93%, 87%, 93%, and 60% among the respective samples (recall).

The results obtained in this project confirm the possibility to use clustering and classification algorithms to categories and predict websites thematic, and consequently, can be used to detect antivaccine websites.

VII. References

Burki, T. (2019). Vaccine misinformation and social media. *The Lancet Digital Health*, *1*(6), e258–e259. https://doi.org/10.1016/S2589-7500(19)30136-0

Centers for Disease Control and Prevention (CDC). (2011). Ten great public health achievements—Worldwide, 2001-2010. *MMWR. Morbidity and Mortality Weekly Report*, *60*(24), 814–818.

Gudivada, V. N., Rao, D. L., & Gudivada, A. R. (2018). Chapter 11 - Information Retrieval: Concepts, Models, and Systems. In V. N. Gudivada & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 38, pp. 331–401). Elsevier. https://doi.org/10.1016/bs.host.2018.07.009

Kari Paul and agencies. (2020, October 13). *Facebook to ban ads discouraging vaccination*. The Guardian. http://www.theguardian.com/technology/2020/oct/13/facebook-vaccine-ads-ban

WHO. (2019a). *Ten threats to global health in 2019*. World Health Organization. https://www.who.int/emergencies/ten-threats-to-global-health-in-2019

WHO. (2019b). *Vaccine Misinformation: Statement by WHO Director-General on Facebook and Instagram*. World Health Organization. https://www.who.int/news-room/detail/04-09-2019-vaccine-misinformation-statement-by-who-director-general-on-facebook-and-instagram