

Deep Learning Techniques in Social Web Text Classification
(Τεχνικές βαθιάς μάθησης στην ταξινόμηση κειμένων του κοινωνικού ιστού)

by

Maria Nefeli Nikiforos (ΠΜΕ201908)
B.Sc., Ionian University, Department of Informatics

A Master's Thesis Submitted in Partial Fulfillment of the
Requirements for the M.Sc.

Research Directions in Information Technology
in the Ionian University, Department of Informatics

© Maria Nefeli Nikiforos, 2021
Ionian University, Department of Informatics

All rights reserved. This dissertation may not be reproduced in whole or in part, by
photocopying or other means, without the permission of the author.

Deep Learning Techniques in Social Web Text Classification
(Τεχνικές βαθιάς μάθησης στην ταξινόμηση κειμένων του κοινωνικού ιστού)

by

Maria Nefeli Nikiforos (ΠΜΕ201908)
B.Sc., Ionian University, Department of Informatics

Supervisory Committee

Associate Professor Katia Lida Kermanidis, Supervisor
(Ionian University, Department of Informatics)

Associate Professor Phivos Mylonas, Departmental Member
(Ionian University, Department of Informatics)

Professor Emmanouil Maragoudakis, Departmental Member
(Ionian University, Department of Informatics)

Supervisory Committee

Associate Professor Katia Lida Kermanidis, Supervisor
(Ionian University, Department of Informatics)

Associate Professor Phivos Mylonas, Departmental Member
(Ionian University, Department of Informatics)

Professor Emmanouil Maragoudakis, Departmental Member
(Ionian University, Department of Informatics)

ABSTRACT

Due to the rapid development of social networks and distribution of vast volumes of user generated content and online information in large amounts, fake news and deception dissemination by malicious accounts is a concerning phenomenon that has emerged in the past decade. In the present Master's thesis, a complete framework for fake news and deception detection from unstructured social web text is proposed, based on machine learning, text mining and NLP techniques and methods. Novel data sets, consisting of tweets containing fake news and real news regarding the Hong Kong protest movement (summer, 2019), are created. Feature sets with both linguistic-oriented and network-oriented features are examined. Binary classification with various machine learning models and different inputs achieves high overall performance compared to recent related work.

Keywords: *Fake News Detection; Deception Detection; Machine Learning; Natural Language Processing; Linguistic Analysis; Text Analysis; Deep Learning; Social Media; Twitter; Pairwise Input*

PUBLICATIONS

The following publications are correlated with segments of the present thesis:

Mouratidis, D., Nikiforos, M. N., & Kermanidis, K. L. (2021). Deep Learning for Fake News Detection in a Pairwise Textual Input Schema (under review). *Computation (MDPI)*.

Nikiforos, M. N., Vergis, S., Styliadou, A., Augoustis, N., Kermanidis, K. L., & Maragoudakis, M. (2020). Fake News Detection Regarding the Hong Kong Events from Tweets (I. Maglogiannis, L. Iliadis, & E. Pimenidis, Eds.). In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations. AIAI 2020 IFIPWG 12.5 International Workshops. AIAI 2020. IFIP Advances in Information and Communication Technology*, Cham. Springer. https://doi.org/10.1007/978-3-030-49190-1_16

Contents

Supervisory Committee	ii
Abstract	iii
Table of Contents	vi
List of Tables	ix
List of Figures	xi
Acknowledgements	xii
Dedication	xiii
1 Introduction	2
1.1 Social Web Text and Twitter	2
1.2 Fake News	3
1.3 Research Questions and Contribution	4
1.4 Agenda	6
2 Related Work	8
2.1 Features in Fake News Detection	8
2.1.1 Surveys	8
2.1.2 Linguistic Analysis & Data Sets	9
2.1.3 Linguistic Analysis & Methods	10
2.1.4 Ensemble & Reinforcement Learning Approaches	10
2.2 Deep Learning in Fake News Detection	11
2.2.1 Surveys	11
2.2.2 Deep Learning Approaches	11
2.2.3 Pairwise Learning Approaches	12

3	Machine Learning	14
3.1	Baseline Classifiers	15
3.1.1	Naive Bayes	15
3.1.2	Random Forest	16
3.1.3	Generalized Linear Model H2O	16
3.2	Deep Learning Models	17
3.2.1	Pairwise Learning	17
3.2.2	Deep Learning H2O	17
4	Methodology	19
4.1	Data	19
4.1.1	Data Set 1	19
4.1.2	Data Set 2	21
4.2	Features	23
4.2.1	Linguistic Features	23
4.2.2	Network Features	24
4.2.3	Feature Selection	25
4.2.4	Feature Encoding	27
5	Experiments	28
5.1	Baseline Classifiers	28
5.1.1	Naive Bayes	28
5.1.2	Random Forest	29
5.1.3	Generalized Linear Model H2O	29
5.2	Deep Learning Models	30
5.2.1	Pairwise Learning	30
5.2.2	Deep Learning H2O	31
6	Results	33
6.1	Baseline Classifiers	33
6.1.1	Naive Bayes	33
6.1.2	Random Forest	35
6.1.3	Generalized Linear Model H2O	37
6.2	Deep Learning Models	39
6.2.1	Pairwise Learning	39
6.2.2	Deep Learning H2O	40

7	Conclusions	43
7.1	Contribution	43
7.2	Future Work	48
A	Abbreviations	51
B	XML: Baseline classifiers experiments (RapidMiner Studio)	52
C	XML: Deep learning models experiments (RapidMiner Studio)	70
	Bibliography	81

List of Tables

Table 4.1 Data Set 2 Categories.	22
Table 4.2 Initial linguistic features.	24
Table 4.3 Initial network features.	25
Table 4.4 Feature Set 1.	26
Table 4.5 Feature Set 2.	26
Table 4.6 Feature Set 3.	27
Table 5.1 Parameters of the pairwise learning model.	31
Table 6.1 Naive Bayes Confusion Matrix. Feature Set 1. Prior to SMOTE.	34
Table 6.2 Naive Bayes Confusion Matrix. Feature Set 1. After SMOTE. .	34
Table 6.3 Naive Bayes Confusion Matrix. Feature Set 3. Prior to and after SMOTE.	35
Table 6.4 Random Forest Confusion Matrix. Feature Set 1. Prior to SMOTE.	35
Table 6.5 Random Forest Confusion Matrix. Feature Set 1. After SMOTE.	36
Table 6.6 Random Forest Confusion Matrix. Feature Set 3. Prior to SMOTE.	36
Table 6.7 Random Forest Confusion Matrix. Feature Set 3. After SMOTE.	36
Table 6.8 GLM H2O Confusion Matrix. Feature Set 1. Prior to SMOTE. .	37
Table 6.9 GLM H2O Confusion Matrix. Feature Set 1. After SMOTE. . .	37
Table 6.10 GLM H2O Confusion Matrix. Feature Set 3. Prior to SMOTE. .	38
Table 6.11 GLM H2O Confusion Matrix. Feature Set 3. After SMOTE. . .	38
Table 6.12 Performance for pairwise learning experiments. Prior to and after SMOTE.	40
Table 6.13 DL H2O Confusion Matrix. Feature Set 1. Prior to and after SMOTE.	41
Table 6.14 DL H2O Confusion Matrix. Feature Set 2. Prior to SMOTE. . .	41
Table 6.15 DL H2O Confusion Matrix. Feature Set 2. After SMOTE. . . .	42
Table 6.16 DL H2O Confusion Matrix. Feature Set 3. Prior to and after SMOTE.	42

Table 7.1	Performance of all models. Feature Set 1. Prior to SMOTE (PS) and after SMOTE (AS). NB: Naive Bayes, RF: Random Forest, GLM: GLM H2O, DL: DL H2O, Acc: Accuracy, Prec: Precision, Rec: Recall, R: Real news, F: Fake news.	47
Table 7.2	Performance of all models. Feature Set 2. Prior to SMOTE (PS) and after SMOTE (AS). PW-1: pairwise learning model (Set 1), PW-2: pairwise learning model (Set 2), DL: DL H2O, Acc: Accuracy, Prec: Precision, Rec: Recall, R: Real news, F: Fake news.	48
Table 7.3	Performance of all models. Feature Set 3. Prior to SMOTE (PS) and after SMOTE (AS). NB: Naive Bayes, RF: Random Forest, GLM: GLM H2O, DL: DL H2O, Acc: Accuracy, Prec: Precision, Rec: Recall, R: Real news, F: Fake news.	48

List of Figures

Figure 5.1 Pairwise learning model: network architecture.	31
Figure B.1 Experiments with Naive Bayes (Feature Set 1, prior to SMOTE). .	53
Figure B.2 Experiments with Naive Bayes (Feature Set 1, after SMOTE). .	55
Figure B.3 Experiments with Naive Bayes (Feature Set 3, prior to SMOTE). .	56
Figure B.4 Experiments with Naive Bayes (Feature Set 3, after SMOTE). .	58
Figure B.5 Experiments with Random Forest (Feature Set 1, prior to SMOTE). .	59
Figure B.6 Experiments with Random Forest (Feature Set 1, after SMOTE). .	61
Figure B.7 Experiments with Random Forest (Feature Set 3, prior to SMOTE). .	62
Figure B.8 Experiments with Random Forest (Feature Set 3, after SMOTE). .	64
Figure B.9 Experiments with GLM H2O (Feature Set 1, prior to SMOTE). .	65
Figure B.10 Experiments with GLM H2O (Feature Set 1, after SMOTE). . .	67
Figure B.11 Experiments with GLM H2O (Feature Set 3, prior to SMOTE). .	68
Figure B.12 Experiments with GLM H2O (Feature Set 3, after SMOTE). . .	69
Figure C.1 Experiments with DL H2O (Feature Set 1, prior to SMOTE). .	71
Figure C.2 Experiments with DL H2O (Feature Set 1, after SMOTE). . . .	72
Figure C.3 Experiments with DL H2O (Feature Set 2, prior to SMOTE). .	74
Figure C.4 Experiments with DL H2O (Feature Set 2, after SMOTE). . . .	76
Figure C.5 Experiments with DL H2O (Feature Set 3, prior to SMOTE). .	78
Figure C.6 Experiments with DL H2O (Feature Set 3, after SMOTE). . . .	80

ACKNOWLEDGEMENTS

Throughout the writing of the present thesis I have received great support and assistance. I would like to thank:

Associate Professor Katia Lida Kermanidis, for her mentoring, support, encouragement, and patience.

Associate Professor Phivos Mylonas & Professor Emmanouil Maragoudakis, for their participation in the supervisory committee.

my colleague Despoina Mouratidis, for her support and our collaboration on projects and publications relevant to my Master's Thesis.

my colleagues Spiridon Vergis, Andreana Styliadou & Nikolaos Augoustis, for our collaboration on projects and publications relevant to my Master's Thesis.

my friends & study group, for supporting me during my M.Sc. studies and the preparation of my Master's Thesis.

my parents, for their advice and constructive criticism.

What are my words?
A little Scheherazade who saves herself
By telling stories.
Raquel Angel-Nagler, Naked Mirrors

Vincit Omnia Veritas.
Latin proverb

Chapter 1

Introduction

1.1 Social Web Text and Twitter

Social web text processing and mining has recently emerged as a challenging task for the Natural Language Processing (NLP), machine learning and data mining research fields. The increasing number of social media accounts connecting through social networks, such as Twitter, Facebook, blogs and wikis, produce a constantly growing volume of written communication through the social web [40]. It has been observed that the performance of NLP tools developed for structured text is significantly reduced when used in unstructured social web text. Consequently, existing methods need to be adapted for social web text, or new methods need to be developed in order to obtain better results and higher performance in unstructured social web text.

Twitter [56] is one of the most popular social networks, with approximately 500 million tweets posted per day. It combines microblogging and social networking. Twitter users can post short messages (tweets), up to 140 characters, to their followers (other Twitter users who have previously connected (social networking) with the followers). Tweets appear publicly on the followees profile page (microblogging).

Social text is typically short in length, low in readability, informal, and syntactically unstructured [39, 40]. It is characterized by great morphological variety and elements of oral speech, while containing misspellings, abbreviations, emojis and emoticons and slang vocabulary. Certain features of tweets are the following: a. tweets are easily distributed to all followers with a retweet, b. the “@” symbol before a username creates a link on the user’s profile page (at-mention), c. the hashtag (#) symbol represents keywords that categorize the tweet on a topic, d. tweet text is short

and abbreviated due to the 140-character limit per post, and e. uppercase letters are used inconsistently.

1.2 Fake News

In the past years, social networks have been rapidly developed, due to the increasing number of users and the dissemination of user generated content and online information in large amounts. This is a phenomenon that has often been exploited by malicious social media accounts (users and entities), which forge, distribute and reproduce fake news and propaganda in vast user networks [1, 2, 3, 4, 17, 18, 20, 28, 30, 35, 36, 40, 50, 53, 58, 59, 61, 63]. As a result, manual fact-checking online information is practically unfeasible.

Fake news is forged information to serve specific interests and ulterior motives. It is disseminated in order to either deceive and make false information believable, or to create doubts about verifiable facts [1, 2, 4, 7, 17, 18, 23, 28, 35, 50, 53, 59]. Another relevant term is propaganda; any information promoting political or other motives and agendas [4, 7, 18, 28, 30, 35, 40, 43, 59, 63].

The language of fake news is deceptive; it intends to provocation and emotional aggravation of the social media users, leading them to reproduce and disseminate the fake information [1, 2, 4, 9, 23, 30, 35, 50], (e.g. “Hong Kong police cannot bow to bad people. The rule of law in Hong Kong does not allow bad people to destroy it.”). Deceptive language often promotes a single viewpoint, and it is highly subjective [4, 23, 30, 43], (e.g. “@ahdee1004 @initiumnews This is the fact. As a Hong Kong person, I have the right to speak for Hong Kong. You are not aware of the actual situation in Hong Kong. As a media person and a Hong Kong person, I have the power to speak the truth on behalf of Hong Kong!”). Grammatical and spelling mistakes, and the use of the same limited set of words are also common features of deceptive language [4, 17, 30, 35]. The development of NLP, text mining and machine learning methodologies and tools aid to provide a more qualitative understanding of both the features of deceptive language (linguistic features), and the account-oriented features of malicious users and entities (network features) [1, 2, 3, 4, 9, 17, 30, 35, 36, 40, 43, 50, 53, 59, 63].

Fake news detection aims at identifying the truthfulness of information by processing the contents and its features [17, 35]. The unstructured and noisy data, the dynamic quality of news and the growing number of social media users, an automated

solution for fake news and deception detection in social networks are considered necessary [1, 3, 4, 7, 9, 28, 30, 40, 43, 50, 53, 59, 61, 63]. Fake news and deception detection on social networks is a quite challenging task and a new research field has emerged, with directions for future work in data-oriented, feature-oriented, model-oriented and application-oriented issues [1, 2, 4, 30, 35, 36, 40, 43, 53, 58, 59, 63].

1.3 Research Questions and Contribution

Due to the rapid development of social networking and distribution of vast volumes of user generated content and online information, fake news and deception dissemination by malicious accounts is a concerning phenomenon that has emerged in the past decade. The issue of fake news and deception detection has been addressed in the past by numerous researchers. However, most proposed approaches rely on a limited set of existing, widely accepted and validated real and fake news data sets. The high performance of previous works in fake news and deception detection relies mainly on using either exclusively account-based features, or exclusively linguistic features. Consequently, the following research questions have emerged:

- **Research Question 1:** Is it possible to design a methodology for creating and validating a real and fake news data set, consisting of social web text?
- **Research Question 2:** Which features can be extracted from a social web text data set? Which (type of) features are more suitable and determinant for efficient fake news and deception detection?
- **Research Question 3:** How is the performance of commonly used machine learning and NLP models for fake news and deception detection affected when the novel data set is used as input?
- **Research Question 4:** How is the performance of pairwise learning models created for different tasks (e.g. machine translation) affected when applied for fake news and deception detection, along with the novel data set as input? What kind of input can be used for creating an efficient pairwise setting?

Significant contribution regarding these research questions is offered in the present thesis, as well as in earlier work conducted by the author [39, 40], published during her

Master’s studies. The results from all these works are gathered and presented extensively. More specifically, the following contribution regarding the research questions is presented:

In earlier work conducted by the author [40]:

- A novel, original, publicly available data set is created [21]. It consists of 2,366 tweets in English, regarding the events in Hong Kong protests of August, 2019. Processing of fake news tweets provided by Twitter, and collecting and validating real news tweets for these events are analyzed thoroughly, proposing a best practice setting for developing fake and real news data sets;
- Multimodal input is used, containing string-based and morphological-textual features (number of syllables, number of long sentences etc.) and higher-level linguistic-oriented features (Flesh-Kincaid level, adverbs-adjectives rate etc.) extracted from the unstructured tweet text, and network-oriented features (Feature Set 1). Additionally, several features are identified as determinant for fake news and deception detection from tweets;
- The performance and obtained results of baseline classifiers are evaluated. These models are commonly used in related work and their performance with the novel data set as input is examined;

In earlier work conducted by the author [39]:

- Tweet vectors are used as input for the pairwise learning models for the first time. Every pair consists of one vector with real tweet data and one vector with real or fake tweet data. The correct classification of the second vector depends on the similarity or diversity when compared to the first vector;
- The input may be either news text or news header. Other researchers have used news articles headers and text as the pair of inputs. However, there is no precedent that categorizes tweets to headers and text based on their linguistic structure. This categorization is outlined in the present thesis, in order to compare the pairwise learning models’ performance with various input types;
- An additional feature set (Feature Set 2) is created, based on the feature set (Feature Set 1) and feature selection performed in the author’s earlier work [40];

- Pairwise learning models are designed in an innovative way, in order to be used for the first time for fake news and deception detection. It combines and fuses multimodal input at various layers, enabling flexibility and high performance.

In the present thesis:

- The approaches regarding data, features, methodology, experiments, results and conclusions in earlier work conducted by the author [39, 40] are gathered and presented as a more complete framework for fake news and deception detection;
- An extensive literature review of the most recent works regarding fake news and deception detection with machine learning and NLP methods is provided. Previous works with different types of data (news articles, tweets, posts, etc.), various categories of features (network, linguistic, etc.), as well as the most efficient network architectures and machine learning models are described in detail;
- The performance and obtained results of the machine learning models of the author's earlier work [40], namely Naive Bayes and Random Forest, are evaluated with the data set (Data Set 1, Feature Set 1) prior to SMOTE over-sampling as input. Their performance with the data set prior to SMOTE as input is examined and compared to their performance of after SMOTE over-sampling;
- An additional feature set (Feature Set 3) is created, based on the feature sets (Feature Set 1 and feature Set 2) and feature selection performed in earlier work conducted by the author [39, 40]. Experiments with different feature sets are conducted and their effect on the machine learning models' performance is examined, for both the baseline classifiers of the author's earlier work [40] and for the additional models proposed in this thesis;
- The performance and obtained results of more machine learning models, namely GLM H2O (baseline classifier) and DL H2O (deep learning) are evaluated. Their performance with the novel data set as input is examined and compared to the performance of the models in the author's previous works [39, 40];

1.4 Agenda

The present thesis is structured as follows:

Chapter 1 describes the features of social web text. It also provides an introduction to fake news and deceptive language, and their dissemination through social networks, such as Twitter. Research questions and contribution are also presented.

Chapter 2 provides an overview of the recent related work regarding fake news and deception detection from social networks, including the most commonly used features and machine learning methodologies for this task.

Chapter 3 presents existing machine learning methodologies and evaluation metrics. Baseline classifiers and deep learning models applied are also described.

Chapter 4 describes the methodology in detail. Creation and preprocessing of the data sets and feature sets are included.

Chapter 5 presents the experimental setups, regarding input, parameters and architecture.

Chapter 6 provides a detailed analysis of the obtained results and achieved performance for each experiment.

Chapter 7 concludes the thesis and states its contribution. The results are discussed extensively. Directions for future work are also provided.

Chapter 2

Related Work

Fake news, spam and propaganda dissemination through social media has caused severe issues, greatly affecting major social and political events. Consequently, literature regarding fake news and deception detection from social networks is increasing, while several researchers attempt to identify the most effective types of data, features and machine learning techniques for such tasks.

In this Section, an overview and detailed analysis of recent related work is presented. Literature regarding the most commonly used features for fake news and deception detection is analyzed in subsection 2.1. Literature regarding deep learning models for fake news and deception detection is presented in subsection 2.2.

2.1 Features in Fake News Detection

2.1.1 Surveys

The work of Vishwakarma & Jain [59] provided a list of the most recent machine learning methods and data sets for fake news and deception detection, according to whether they are applied to textual or image news data. Merryton & Augusta [36] created a survey for identifying the most efficient baseline classifiers and deep learning models for fake and spam messages detection. Additionally, they discussed the most common NLP preprocessing methods and tools, and finally they described the existing linguistic feature sets and data sets.

Han & Mehta [20] defined the fake news types and main linguistic features. They also provided an evaluation and comparison of the performance of baseline classifiers and the performance of deep learning models on fake news and deception detection,

considering accuracy and lightweights. The survey of Shu et al. [53] presented a set of all existing definitions of fake news and the differences between different features. It also explored the effect of fake news dissemination on both social and traditional media, and analyzed the most common fake news detection approaches.

2.1.2 Linguistic Analysis & Data Sets

The significance of textual and linguistic features for fake news and deception detection is explored in several works, including the earlier work conducted by the author [40]. Zervopoulos et al. [63] provided a novel data set regarding the protest movement of summer, 2019 in Hong Kong; fake news: 3,908 tweets in English, and Chinese translated into English (Google Translation API), and real news: 5,388 tweets in English from both acknowledged news agencies and individuals (e.g. journalists). Their feature set consisted of exclusively linguistic features, while their goal was to identify linguistically similar tweets. Data preprocessing and experiments were conducted with Python (Scikit-Learn, NLTK). Experiments with Naive Bayes, SVM, C4.5 and Random Forest achieved an average of 92.1% F1-score, Random Forest being the most efficient classifier. The approach of Jeronimo et al. [23] proposed a new data set; fake news: 95 news of 2 fact-checking services (2010-2017, Brazil), and real news: 207,914 news articles of 2 mainstream media platforms (2014-2017, Brazil), in the columns Politics, Sports, Economy and Culture. The feature set contained the semantic distance between the data and 5 subjectivity lexicons (argumentation, presupposition, sentiment, valuation, and modalization). Experiments were conducted with XGBoost, Random Forest (along with Bag of Words and TF-IDF) being the most efficient in inter domain cases.

Wang et al. [60] created LIAR, a publicly available data set for fake news and deception detection, consisting of 12,800 manually labeled short statements regarding several different topics from PolitiFact. Surface-level linguistic patterns were explored. Experiments were conducted with hybrid Convolutional Neural Network (CNN), setting a benchmark for fake news and deception detection on the proposed data set. Finally, Shu et al. [52] created FakeNewsNet, a new fake news data repository, containing 2 data sets with numerous features regarding news content, social context, and spatio-temporal information.

2.1.3 Linguistic Analysis & Methods

Other works explored the application of linguistic analysis to machine learning methods for fake news and deception detection, as well as for automated fact-checking. Houvardas & Stamatatos [22] focused on fact-checking information through developing an automatic authorship identification tool, which is considered as a preliminary deception detector, based mostly on text analysis. In the work of Mahyoob et al. [35] 20 PolitiFact posts represented real news and 20 Facebook posts represented fake news, resulting in 6 classes. Both a qualitative and a quantitative data analysis were applied with the QDA tool, by comparing the posts regarding their linguistic and textual features. Shojaee et al. [51] attempted fake news and deception detection through reviews, by extracting lexical and syntactic features. Experiments with SVM and Naive Bayes models were conducted, achieving 84% and 74% Accuracy, respectively.

Alowibdi et al. [5] concluded that a Naive Bayes-Decision Tree hybrid was the most efficient model for fake news and deception detection with network features, achieving 85% Accuracy. Songram et al. [54] conducted experiments with KNN and SVM to identify deceptive messages in Thai social media. SVM was the most efficient, achieving 99.21% Accuracy. Other works are those of Pérez-Rosas & Mihalcea [44], conducting Linguistic Inquiry and Word Count (LIWC) word class analysis for deceptive text identification, and Hai et al. [19], using logistic regression in a semi-supervised learning setting to detect fake news and deceptive language.

2.1.4 Ensemble & Reinforcement Learning Approaches

Certain studies designed and applied ensemble and reinforcement learning techniques with high efficiency in fake news and deception detection. In a study by Agarwal & Dixit [2], the LIAR data set was used for representing fake news, and a Kaggle data set (20,801 USA news reports) was used for representing the real news, as input for a binary classification schema. Credibility scores and various textual and linguistic features were extracted, while data were normalized and tokenized. Data preprocessing and experiments were conducted with Python-based tools and libraries, such as Scikit-Learn, pandas, numpy, Keras, and NLTK. The produced ensemble consisted of a Support Vector Machine (SVM), a CNN, a Long Short-Term Memory (LSTM), a k-Nearest Neighbor (KNN), and a Naive Bayes classifier, along with Bag of Words, Term Frequency–Inverse Document Frequency (TF-IDF) and n-grams. The best re-

sults were obtained with the LSTM (97% Accuracy). Wang et al. [61] developed WeFEND, a framework for automatic annotation of news articles. It used user reports from WeChat to weakly supervise fake news and detection detection. Textual and linguistic features were extracted. Experiments with reinforcement learning, using LIWC and LSTM, achieved up to 82% Accuracy.

2.2 Deep Learning in Fake News Detection

2.2.1 Surveys

Perera [43] reviewed existing deep learning models and techniques for both manual and automatic fake news detection, and defined 7 distinct levels of fake news, considering their context and the motive for their composition and dissemination. Finally, they presented different methods of processing of fake news by algorithms, specially implemented in social media platforms. Alam & Ravshanbekov [4] also defined fake news and discussed the positive impact of combining NLP and deep learning for automatic fake news and deception detection. Ruchansky et al. [47] designed and implemented a hybrid deep learning model for fake news and deception detection. Their data set consisted of 992 tweets, 233,719 users, and 592,391 interactions. Their feature set consisted of information regarding both text and user behavior. Experiments achieved up to 82.9% Accuracy.

2.2.2 Deep Learning Approaches

Several researchers proposed and implemented deep learning models and techniques, which proved to be highly efficient in terms of Accuracy. Sansonetti et al. [50] presented a new data set (568,315 tweets). The tweets were matched with various news indexed in PolitiFact (62,367 news: 34,429 fake news, and 29,938 real news), and 4,022 user profiles (2,013 who publish mainly fake news, 2,008 who publish mainly real news). Their feature set included both network and linguistic features. Experiments for offline and online analysis were conducted with CNN, LSTM, dense layer and baseline classifiers (SVM and KNN), reaching up to 92% Accuracy. In the work of Kumar et al. [30], a comparison among several ensembles was performed, regarding binary classification on 1,356 news (Twitter) and 1,056 real and fake news (PolitiFact). A data set per topic was produced, which was afterwards tokenized and encoded

with BeautifulSoup, Python, GloVe and GPy. Experiments were conducted with embeddings, CNN and LSTM (both ensemble and bidirectional). The CNN and bidirectional LSTM ensembled network with attention mechanism obtained the best results (88.78% Accuracy).

Another approach is that of Alves et al. [7], who created a novel, binary class, data set. It included 2,996 articles in Brazilian Portuguese (May-September, 2018). It was normalized and tokenized with Keras and TensorFlow. The experiments were conducted with a bidirectional and a regular LSTM, and a dense layer. The three-layer deep bidirectional LSTM with trainable word embeddings was the most efficient, with up to 80% Accuracy. Victor [58] used the PHEME data set and the LIAR data set as input, and conducted experiments with a deep two-path CNN and a bi-directional RNN for supervised and unsupervised learning, reaching up to 83% Accuracy. Finally, Koirala [28] proposed a new data set, consisting of 4,072 news articles (Webhose.io), to detect fake news regarding COVID-19. They used linguistic features and conducted experiments with baseline classifiers, LSTM and dense layer, with Accuracy from 70% to 80%.

2.2.3 Pairwise Learning Approaches

Pairwise learning models are greatly used for machine learning tasks, including fake news and deception detection. In pairwise frameworks, the training data set consists of lists of items that are specifically ordered within each list. Koppel et al. [29] proposed a straightforward pairwise learning model for ranking tasks. Experiments with the LETOR MSLR-WEB10K, MQ2007 and MQ2008 data sets were conducted with Tensorflow and the Adam-Optimizer. In the work of Dong et al. [17], the PHEME data set was used as input for semi-supervised binary classification with baseline classifiers, LSTM and a deep two-path learning model containing three CNNs. Their efficiency was higher, compared to the supervised learning models, especially when the distribution between the training and test data sets was unbalanced. They also proved to be more resistant and robust considering the issue of overfitting.

Agrawal et al. [3] worked with tweets that consisted of mainly multimedia content. The training set consisted of 5,000 real and 7,000 fake news. The test set consisted of 1,200 real and 2,500 fake news. The novelty of their approach lies on the fusion of a pairwise ranking approach with a classification system. Their feature set consisted of image-based features, Twitter user-based features and tweet-based

features. For classification, a deep neural network, logistic regression and SVM were used, combined with n-grams and doc2vec vectors. The ranking was calculated based on the distance between the features (contextual comparison) of tweets of the same topic (defined by hashtag). The ranking system outputs were incorporated within the classification system. Up to 89% and 78% Accuracy was reached for real and fake news, respectively. The approach of Bahad et al. [9] included two unstructured news data sets (open Machine Learning Repository, Kaggle) for binary classification. Experiments were conducted with LSTM, RNN and CNN, implemented with Python and TensorFlow. The bi-directional LSTM-RNN had the best performance (up to 98% Accuracy). Abdullah et al. [1] tokenized news of 12 different topics. In their work, the prediction of the topic determines the fake from the real news (12 classes). Experiments were conducted on Kaggle’s cloud, with CNN, LSTM and dense layer (up to 97.5% Accuracy).

In a quite different pairwise machine learning setting, Mouratidis et al. [38] proposed a general deep learning model architecture for classifying parallel translations of 2 machine translation model outputs and 1 human (reference) translation, with linguistic information. Their schema achieved the best evaluation score when information from embeddings and simple features were used for small data sets. Augenstein et al. [8] suggested a framework which uses information from embeddings in a multi-task learning experiment. They evaluated their approach on numerous parallel classification tasks for sentiment analysis. Their conclusions indicated that when the learning framework utilized the ranker scores, the classification system managed to outperform a simple classification system.

Chapter 3

Machine Learning

“Machine Learning is programming computers to optimize a performance criterion using example data or past experience” [6]. Machine learning techniques, along with NLP and text mining, have significant applications in fake news and deception detection from social text data, due to their ability to adapt to and handle large data sets quite efficiently. In most machine learning tasks, a process begins with a general model which can perform certain tasks, depending on the values of specified parameters; during the training phase, these values are adjusted according to the training data, in order to adapt the model to the specific data and target task; during the testing phase, the adapted model is tested with new, previously unknown data, while its performance regarding the target task is evaluated [6, 48]. Machine learning may be supervised, unsupervised, semi-supervised, or reinforcement.

In supervised learning, the aim is to learn a mapping from the input to an output whose correct values are previously known and provided by a supervisor (labeled data) [6, 48]. Each learning example of the training set consists of a vector of features and a desired output value, commonly referred to as label or class. In this thesis, the output receives distinct values, and consequently the problem to which supervised machine learning algorithms are applied is classification. In unsupervised learning, the correct values of the output are neither known beforehand, nor provided by a supervisor (unlabeled/unclassified data) and thus the aim is to find the regularities in the input and detect the underlying structure of the data [6, 48]. Semi-supervised learning combines both supervised and unsupervised learning; algorithms are trained with both labeled and unlabeled data [6]. In reinforcement learning, an agent learns behavior through a trial-and-error process in a dynamic environment [25], and it is applied when the target task results from a sequence of actions [6]. The actions to

take in each step are not known; several actions are tried dynamically to identify which of them are more efficient [37, 48, 57].

There are several evaluation metrics to evaluate the performance of machine learning algorithms [24, 55]. Accuracy is calculated to show the correctness of the prediction for the data examples of all classes. Precision describes the ratio of total classifications made by the algorithm on the test set that are correct. Recall is the ratio of the total test data that are correctly identified by the classifier. F1-measure, also known as F1 score, is the harmonic mean of precision and recall.

Mean Squared Error (MSE) measures the average of all the average squared differences between true and predicted values. Specifically, it measures the average squared deviation between a predicted label and the input's true label. Related to MSE, Root Mean Squared Error (RMSE) is the square root of the average of squared error, and focuses mainly on large errors. Coefficient of determination (R^2) measures the machine learning model's ability to replicate the true labels, based on the proportion of total variation of labels.

In the following subsections, baseline classifiers (Naive Bayes, Random Forest, and GLM H2O) and deep learning models (pairwise learning model and DL H2O model) used for the machine learning experiments of the present thesis and in earlier work conducted by the author [39, 40] are defined and their implementation is described.

3.1 Baseline Classifiers

Certain classic and simple classifiers are commonly used for experiments, in order to establish a baseline for the achieved performance. The machine learning algorithms which are used for the experiments with baseline classifiers (Section 5 and earlier work conducted by the author [40]) are described in the following subsections.

3.1.1 Naive Bayes

Bayesian inference is a statistical method that uses Bayes' theorem to redefine the probability of a hypothesis as more information or data become available [15]. The supervised machine learning algorithm Naive Bayes is characterized with high-bias and low-variance. It is able to generate an efficient classification model, even with a small data set and low computational complexity. Its application is common in text categorization, spam and bot detection, sentiment analysis and recommender

systems.

The “naive” assumption this algorithm is based on is that, given the value of the label/class, the value of any feature is independent of the value of any other feature. This independence assumption has a great impact on the simplification of the calculations required to produce the probability model. However, a weakness of this algorithm emerges: if within the training data a given value of a feature never appears within a given class, then the conditional probability is set to zero. When this zero value is multiplied with the other probabilities, these values become also zero and thus the results may be misleading. This weakness is addressed with the Laplace correction technique, where value one is added to each measurement to avoid the occurrence of zero values, while the effect on the estimated probabilities is negligible in most cases.

3.1.2 Random Forest

Random Forest is an ensemble machine learning method for classification [16]. During the training phase, multiple random decision trees are constructed and trained on bootstrapped subsets of the training set. Each of their nodes represents a splitting rule for one specific feature; a subset of features is considered for the splitting rule selection. New nodes are constructed repeatedly, until the stopping criteria are met. During the testing phase, each random decision tree produces a prediction for each test set example by following the branches of the tree (according to the splitting rules) and evaluating the leaf (prediction). The majority of examples define the class predictions; the estimations are obtained through the average of values reaching a leaf. This process results in a voting model; all single predictions are considered of equal significance and they are based on subsets of examples.

Pruning is a technique that may be used to reduce the complexity of the model. It replaces sub-trees with limited predictive power with leaves. This technique may be applied either during the generation of the random decision trees (prepruning), or after. Random Forest is more robust regarding overfitting, compared to decision trees.

3.1.3 Generalized Linear Model H2O

Generalized Linear Model H2O (GLM H2O) is an extension of the traditional linear models, which fits generalized linear models to the data, by maximizing the log-

likelihood [14]. Its strength rely on the parallel and fast model fitting computation. It also scales extremely well for a limited number of predictors with non-zero coefficients. GLM H2O algorithm is implemented in RapidMiner Studio with H2O 3.30.0.1. It uses a 1-node local H2O cluster to run the algorithm on it.

Certain parameters of the model affect the achieved performance. The family parameter adapts the model according to the type of input data (numeric, polynomial, etc.). Various solvers are provided to adapt the model according to the number and context of features. The level of parallelism is determined by adjusting the number of threads. Finally, certain parameters for regularization, handling of missing values, and early stopping are also available.

3.2 Deep Learning Models

More sophisticated models are applied to improve the overall performance, as well as to discover more efficient approaches. The models which are used for the experiments with deep learning (Section 5 and in earlier work conducted by the author [38]) are described in the following subsection.

3.2.1 Pairwise Learning

Pairwise learning models are commonly used for various machine learning tasks and applications, including fake news and deception detection. In pairwise frameworks, the training data set consists of lists of items that are specifically ranked within each list [29]. At the training stage, the model evaluates (compares) a pair of inputs and computes a preference score, eventually determining the one that is the most relevant (ranking) [3]. Consequently, pairwise learning frameworks may be applied in binary classification tasks.

3.2.2 Deep Learning H2O

Deep Learning H2O (DL H2O) implements a multi-layer feed-forward artificial neural network [13]. It is trained with stochastic gradient descent using back-propagation, containing hidden layers of neurons with various activation functions. DL H2O algorithm is implemented in RapidMiner Studio with H2O 3.30.0.1. It uses a 1-node local H2O cluster to run the algorithm on it (parallel execution). Each node trains a copy

of the global model parameters on its local data with asynchronous multi-threading, periodically contributing to the global model via model averaging across the neural network.

Activation function may be Tanh (scaled and shifted sigmoid), Rectifier, Maxout, and Exponential Rectifier. Several parameters, such as adaptive learning rate, rate annealing, momentum training, dropout, L1/L2 regularization, early stopping and handling of missing values, are also feasible.

Chapter 4

Methodology

In this Section, the methodology regarding data and features is presented. More specifically, data collection, annotation, preprocessing, normalization, encoding and feature extraction for both types of experiments are described extensively.

4.1 Data

The collection of the data used for the creation of both Data Set 1 and Data Set 2 was from Twitter Safety [49], (fake news data) and was performed with a crawler script in Python the Twitter Standard Search API [31], (real news data). The collected tweets, 2,366 in total, are written in English, their content is relevant to the Hong Kong events and protest movement, and were posted in the summer of 2019. Data Set 1 and Data Set 2 were created in the context of earlier work conducted by the author [40] and [39], respectively. Network features (Table 4.3) which were available along with these tweets were also collected. Linguistic features were not available, and thus were extracted from the text of each tweet, as described in subsection 4.2.

4.1.1 Data Set 1

Data Set 1 is used as input for the machine learning experiments with baseline classifiers and DL H2O from the deep learning models. In this subsection, the creation, preprocessing and preparation of Data Set 1 is described.

Fake News

In the summer of 2019, Twitter Safety [49] disclosed 936 accounts from People’s Republic of China, which were deliberately attempting to sow political discord in Hong Kong, aiming to thwart the legitimacy, and political convictions and notions of the Hong Kong protest movement. These accounts posted and spread false information systematically, and consequently were suspended due to violations of Twitter’s manipulation policies (spam, coordinated activity, fake accounts, featured activity, and ban evasion).

The fake news tweets of Data Set 1 were collected from these Twitter accounts. 1,703,470 tweets were preprocessed, in order to extract tweets that actually: a. contain text, b. are written in English, and c. include content relevant to the Hong Kong events and protest movement (summer of 2019). A label feature was added for each tweet, with value “fake” (label/class). After preprocessing, the final fake news data set consists of 272 tweets.

Real News

In order to represent and include tweets containing true and valid information in Data Set 1, 9 Twitter accounts of renowned news agencies were selected, namely: BBC Asia, BBC News (World), CCTV, China Daily, China Xinhua News, China.org.cn, Global Times, People’s Daily (China), and SHINE. As a result, tweets posted by these accounts from summer to winter of 2019 were accessed and collected.

2,133 tweets were preprocessed, in order to extract tweets that actually: a. contain text, b. are written in English, and c. include content relevant to the Hong Kong events and protest movement (summer of 2019). A label feature was added for each tweet, with value “real” (label/class). After preprocessing, the final real news data set consists of 2,094 tweets.

Data Preparation

Imbalance between the number of fake and real news learning examples (272 to 2,094 learning examples, respectively) was observed. This unequal representation may affect negatively the performance of the minority class (label feature “fake”) prediction. Therefore, over-sampling of the minority examples was considered necessary. Consequently, experiments were conducted with Data Set1 prior to SMOTE and after SMOTE. SMOTE over-sampling was performed with the RapidMiner Studio version

9.5 [45], as defined in [11], and “synthetic” fake news training examples were created. According to Chawla et al. [11], it needs to be ensured that the test set for the machine learning experiments does not include any “synthetic” examples. Consequently, prior to SMOTE over-sampling, Data Set 1 was split in 80% training set and 20% test set, with stratified sampling; it preserves the label distribution in the subsets in order to be the same proportionally as in the whole data set. After the split, the training set consists of 1,893 examples: 1,675 with value “real” and 218 with value “fake” for the label feature, while the final test set consists of 473 examples: 419 with value “real” and 54 with value “fake” for the label feature. Test set remained pure. Finally, SMOTE over-sampling was applied to the training set, resulting in a training set consisting of 3,350 examples: 1,675 with value “real” and 1,675 (of which the 1,457 are “synthetic” examples) with value “fake” for the label feature.

Feature selection was also applied (described in subsection 4.2), in order to identify the features which are more suitable and significant for fake news detection. This task was performed with the RapidMiner Studio version 9.5 [45].

4.1.2 Data Set 2

Data Set 2 is used as input for the machine learning experiments with pairwise learning models from the deep learning models. In this subsection, the creation, preprocessing and preparation of Data Set 2 is described.

Origins

Data Set 2 was based on Data Set 1. It consists of 2,363 tweets (272 fake news and 2,092 real news) written in English, and their content is relevant to the Hong Kong events and protest movement of August, 2019. Data Set 2 contains the tweet text that is used as an input to the neural networks of deep learning models described in Section 5, subsection 5.2.

Categorization

In order to be used as input in the pairwise learning schema, the tweets of Data Set 2 were categorized into 4 distinct categories, depending on both the class (real or fake) and the type of the tweet text (header or text).

Tweets that make a single-sentence statement, e.g. “Bricks on the street block Hong Kong’s democracy.”, are considered as headers (real or fake, depending on the

class). Their form is quite similar to newspaper headlines. Tweets that are longer than one sentence, e.g. “You thought this is on behalf of the people in Hong Kong. On the contrary, it is a rascality of putting the “false freedom” label on the will of most of Hong Kong people.”, are considered as text (real or fake, depending on the class). Their form is quite similar to newspaper articles’ main text. Consequently, the 4 categories (sub-data sets) are: a. real header, b. real text, c. fake header, and d. fake text. In Table 4.1, more information regarding the categories and their respective sub-data sets is provided.

Table 4.1: Data Set 2 Categories.

Category	Number of Tweets
Real Header	1.027
Fake Header	127
Real Text	1.065
Fake text	144

Data Preparation

The pairwise learning models, described in Section 5, subsection 5.2, require two inputs. Consequently, the two tweet text inputs were defined as $T1$ and $T2$. For the first set of experiments, $T1$ belongs to the real header category and $T2$ can either belong to one of the following categories: real text, fake text, fake header. For the second set of experiments, $T1$ belongs to the real text category and $T2$ can either belong to one of the following categories: real header, fake text, fake header.

Due to class imbalance, with fake news being the minority class, SMOTE over-sampling was applied. After SMOTE over-sampling [11], the total number of tweets increased from 2,363 to 3,766. According to Chawla et al. [11], it needs to be ensured that the test set for the machine learning experiments does not include any “synthetic” examples. Consequently, prior to SMOTE over-sampling, Data Set 2 was split in 80% training set and 20% test set, with stratified sampling; it preserves the label distribution in the subsets in order to be the same proportionally as in the whole data set. After the split, the training set consists of 1,891 examples: 1,674 with value “real” and 217 with value “fake” for the label feature, while the final test set consists of 472 examples: 418 with value “real” and 54 with value “fake” for the label feature. Test set remained pure. Finally, SMOTE over-sampling was applied to the training set, resulting in a training set consisting of 3,294 examples: 1,674 with value

“real” and 1,620 (of which the 1,403 are “synthetic” examples) with value “fake” for the label feature. Both sets of experiments were conducted with Data Set 2 prior to SMOTE and after SMOTE over-sampling, in order to examine whether it improves the obtained results and performance of the model.

Feature selection was also applied (described in subsection 4.2), in order to identify the features which are more suitable and significant for fake news detection.

4.2 Features

The collected tweet texts contain numerous features, both linguistic-oriented (Table 4.2) and network-oriented (Table 4.3). Following feature selection and encoding, 3 feature sets (Table 4.4, Table 4.5 and Table 4.6) are described in this subsection.

4.2.1 Linguistic Features

Several linguistic features were extracted from the tweet text to depict its specific language traits and forms. They were selected according to Burgoon et al. [10] and Li et al. [32], who applied linguistic analysis for fake news and deception detection, as well as attitude identification. Linguistic features provide significant information for the tweets’ authors, regarding their motives and intentions, the way they express themselves, including their opinions, sentiment and emotions, as well as their educational level and background in general.

The extraction of linguistic features was performed with a python script. The Natural Language Toolkit (NLTK) library [12] was used to tag the tweets by Part-Of-Speech (POS), in order to identify adjectives and adverbs in the text. The threshold for the long and short sentences was set according to Pennebaker et al. [42]; sentences with less than 8 words were considered as “short”, while sentences with more than 25 words were considered as “long”. Similarly, words with more than 6 letters were considered as “big”. Flesh-Kincaid level [62], a unique feature, indicates how difficult it is to understand a specific piece of text; high Flesh-Kincaid level means highly understandable text, and vice versa.

The average number of syllables per word is one of the most commonly used linguistic features in the literature. This feature was extracted with the Loughran Master Dictionary [33, 34]; it determined which tokens of the tweet text correspond to proper English words, and then it produced their syllable count. The tokens were

matched with the words of the dictionary by ignoring punctuation marks, hashtags, at-mentions, and hyperlinks. Another challenge was that the tweet text usually contains tokens that are either misspelled, or represent a combination of words, e.g. #HongKongProtest. In these cases, the average syllable count of English words was assigned by default as the token’s syllable count.

The linguistic feature set (Table 4.2) consists of the following features: a. tweet text, b. number of syllables, c. number of words, d. number of sentences, e. number of big words, f. number of syllables per word, g. number of short sentences, h. number of long sentences, i. Flesh-Kincaid level, j. average number of words per sentence, and k. rate of adjectives and adverbs.

Table 4.2: Initial linguistic features.

Feature	Tweet Example
Tweet text	While Hong Kong police are dealing with violent acts ranging from vandalizing shops, setting fires and assaulting
Num syllables	35
Avg syllables	1.8
Avg Words in Sentence	18
Flesh-Kincaid	13.55
Num big Words	8
Num long sentences	0
Num short sentences	1
Num sentences	0
Num words	19
Rate adverbs adjectives	0.0526

4.2.2 Network Features

Network features were collected directly from Twitter, at the same time with the corresponding tweets, to access information regarding the account that posted each tweet and its connections throughout Twitter. They were selected according to Shu et al. [52], who presented a fake news data repository, consisting of data collected from Twitter, and Shu et al. [53], who provided a survey, reviewing state-of-the-art fake news detection approaches on social media. Network features display the structure of user (social) networks and the interactions within them, connected with common interests, topics and relations. Interactions within social networks enable information

diffusion to a vast number of users; fake news dissemination processes tend to form an echo chamber cycle. Consequently, using network-oriented features to represent types of network patterns is significant for fake news detection.

The network feature set (Table 4.3) consists of the following features: a. user id, b. user display name, c. user screen name, d. follower count, e. following count, f. account creation date, g. tweet time, h. in reply to user id (reply to specific user), i. like count, j. retweet count, and k. number of URLs.

Table 4.3: Initial network features.

Feature	Tweet Example
User id	487118986
User display name	China Xinhua News
User screen name	XHNews
Follower count	12616791
Following count	61
Account creation date	12/3/10 12:00 AM
Tweet time	12/17/19 2:34 PM
In reply to user id	2147483647
Like count	37
Retweet count	12
Num URLs	1

4.2.3 Feature Selection

Regarding Data Set 1, to create the final feature set, feature selection was performed with the RapidMiner Studio version 9.5 [45]. The most suitable and informative features for fake news and deception detection were selected. Specific network features were observed to be determinant for fake news and deception detection (most to least determinant): a. user id, b. account creation date, c. following count, d. user display name, and e. user screen name. During the initial experiments, it was observed that machine learning algorithms take into consideration only one of these features (from most to least determinant), and ignore completely every other feature (linguistic or network). Consequently, to ensure that the models actually do “learn”, and to obtain reliable results from the experiments, these features were removed from the feature set. As a result, machine learning experiments with baseline classifiers (Section 5, subsection 5.1) were conducted with the training and test sets that were designed as described in subsection 4.1.1, and a feature set (Feature Set 1) consisting of the 18

remaining features, as shown in Table 4.4.

Table 4.4: Feature Set 1.

Linguistic features	Network features
Tweet text	Follower count
Num syllables	Num URLs
Avg syllables	Tweet time
Avg Words in Sentence	In reply to user id
Flesh-Kincaid	Like count
Num big Words	Retweet count
Num long sentences	
Num short sentences	
Num sentences	
Num words	
Rate adverbs adjectives	

Regarding Data Set 2, to create the final feature set, feature selection was performed. The following network features were removed from the feature set, due to the large number of missing values: a. user display name, b. user screen name, and c. in reply to user id. Additionally, tweet text was used as input and not as a feature, in contrast with Feature Set 1. As a result, machine learning experiments with the pairwise learning model (Section 5, subsection 5.2) were conducted with the training and test sets that were designed as described in subsection 4.1.2, and a feature set (Feature Set 2) consisting of the 18 remaining features, as shown in Table 4.5.

Table 4.5: Feature Set 2.

Linguistic features	Network features
Rate adverbs adjectives	Follower count
Num syllables	Account creation date
Avg syllables	Tweet time
Avg Words in Sentence	Following count
Flesh-Kincaid	Like count
Num big Words	Retweet count
Num long sentences	Num URLs
Num short sentences	User id
Num sentences	
Num words	

For the experiments with the DL H2O model, Data Set 1 was used. In order to compare the results of this model to those of the baseline classifiers and the pairwise

learning model, 3 sets of experiments were conducted with: a. Feature Set 1, b. Feature Set 2, and c. a feature set (Feature Set 3) including only the features that are common in both Feature Set 1 and Feature Set 2. To create Feature Set 3, feature selection was performed with the RapidMiner Studio version 9.8 [45]. Feature Set 3 consists of 16 features, as shown in Table 4.6.

Table 4.6: Feature Set 3.

Linguistic features	Network features
Tweet text	Follower count
Num syllables	Num URLs
Avg syllables	Tweet time
Avg Words in Sentence	Retweet count
Flesh-Kincaid	Like count
Num big Words	
Num long sentences	
Num short sentences	
Num sentences	
Num words	
Rate adverbs adjectives	

4.2.4 Feature Encoding

Feature encoding was considered necessary for Feature Set 2, in order to be used by the pairwise learning models. More specifically, for the network feature account creation date, all dates were converted from text to numerical, while for the network feature tweet time, the dates were converted from text to numerical and the times were converted from a 12-hour time to a 24-hour time to escape the ambiguity between p.m. and a.m. hour notations. Finally, all features were scaled with the MaxAbsScaler, as described in Pedregosa et al. [41].

Chapter 5

Experiments

In this Section, a description of the experimental setups is provided. The selected baseline classifiers, their parameters and input, as well as the proposed deep learning models their network architectures, their parameters and input are presented.

5.1 Baseline Classifiers

Several experiments with baseline classifiers were conducted. The tool used for the experiments with baseline classifiers, the collection of their results, the comparison of the machine learning models, and the identification of wrong predictions is the RapidMiner Studio version 9.5 [40] and version 9.8.

5.1.1 Naive Bayes

The first set of experiments was conducted with the Naive Bayes classifier, since it is commonly used in previous works [2, 5, 51, 63]. Data Set 1 was used as input; the classifier was at first trained with the SMOTE over-sampled training set and tested with the original test set (as described in Section 4, subsection 4.1), and then it was trained with the original (prior to SMOTE over-sampling) training set and tested with the original test set, in order to examine whether and how SMOTE over-sampling affects the performance. Additionally, these experiments were conducted with Feature Set 1 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance. Laplace correction was used as a parameter, in order to smooth the conditional probabilities. A detailed description of this parameter is provided in Section 3, subsection 3.1.

5.1.2 Random Forest

The second set of experiments was conducted with the Random Forest classifier, since it is commonly used in previous works [23, 46, 63]. The Random Forest algorithm aims to minimize the overall error rate. Consequently, for unbalanced data sets, the majority class will get a relatively lower error rate while the minority class will have a larger error rate [46]. Data Set 1 was used as input; the classifier was at first trained with the SMOTE over-sampled training set and tested with the original test set (as described in Section 4, subsection 4.1), and then it was trained with the original (prior to SMOTE over-sampling) training set and tested with the original test set, in order to examine whether and how SMOTE over-sampling affects the performance. Additionally, these experiments were conducted with Feature Set 1 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance. The following parameters were defined: a. the gain ratio was defined as the split criterion (no pruning), with maximal depth defined to 50, b. the number of trees to generate was set to 100, c. the subset ratio of randomly chosen features to test was set to 2, and d. the majority vote was defined as the voting strategy; selection of the label that is predicted by the majority of the tree models. A detailed description of these parameters is provided in Section 3, subsection 3.1.

5.1.3 Generalized Linear Model H2O

Unlike previous works, the third set of experiments was conducted with the GLM H2O classifier. It is the first time (to the best of the author’s knowledge) this classifier is applied for fake news and deception detection. Data Set 1 was used as input; the classifier was at first trained with the SMOTE over-sampled training set and tested with the original test set (as described in Section 4, subsection 4.1), and then it was trained with the original (prior to SMOTE over-sampling) training set and tested with the original test set, in order to examine whether and how SMOTE over-sampling affects the performance. Additionally, these experiments were conducted with Feature Set 1 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance. The following parameters were defined: a. the family, defined by automatic selection, b. the solver, defined by automatic selection, and c. regularization and standardization with default values. A detailed description of these parameters is provided in Section 3, subsection 3.1.

5.2 Deep Learning Models

Several experiments with deep learning were conducted. Experiments with pairwise learning models, the collection of their results, the comparison of the machine learning models, and the identification of wrong predictions were conducted with Python (pandas, numpy and Keras [26]), [39]. The tool used for the experiments with DL H2O, the collection of their results, the comparison of the machine learning models, and the identification of wrong predictions is the RapidMiner Studio version 9.8.

5.2.1 Pairwise Learning

Experiments were conducted with the pairwise learning model, since pairwise learning models and similar network architectures are commonly used in previous works [1, 3, 8, 9, 17, 29, 38].

The proposed pairwise ranking approach considered fake news and deception detection task as a binary classification task. A pair of tweets, $T1$ and $T2$, consist the input. An automatically calculated (with the Keras library [26]) embedding layer was used for $T1$ and $T2$ of each pair of tweet text input. The embedding layer consists of 18 nodes. Regarding the input dimensions of the embedding layer, the vocabulary of each input tweet text was taken into consideration (most frequent words). One-hot function was applied for encoding. The annotation for the binary classification task is calculated according to the following:

$$y = \begin{cases} 0, & \text{if } T1 \text{ is a real tweet and } T2 \text{ is a real tweet} \\ 1, & \text{if } T1 \text{ is a real tweet and } T2 \text{ is a fake tweet} \end{cases}$$

where y is the classification label. Based on $T1$ and $T2$, the embedding vectors $EmbT1$ and $EmbT2$ were created on the embedding layer and were preprocessed with the MaxAbsScaler. At the next stage, the embedding vectors were integrated in a parallel setting and became the input to the hidden layer. The output of the hidden layer becomes the input to the last layer. There, a matrix $F[i,j]$ was added; a two-dimension matrix containing linguistic and network features. The output label is a random variable (calculated with maximum likelihood estimation), to minimize the discrepancy between the predicted and the true classification labels. The classification task was modeled as a Bernoulli distribution.

The model architecture for all of the experiments was defined as shown in Figure 5.1. Dense layers 1 & 2 contained 128 Hidden Units, and dense layer 3 contained 1

Hidden Unit (last layer). The output layer used Sigmoid as the activation function, while dense layers 1 & 2 used Relu, and dense layer 3 used Sigmoid as the activation functions. The dropout of dense layers was set to 0.4. All sets of experiments were conducted with 10-fold cross validation. More parameters of the pairwise learning model are shown in Table 5.1.

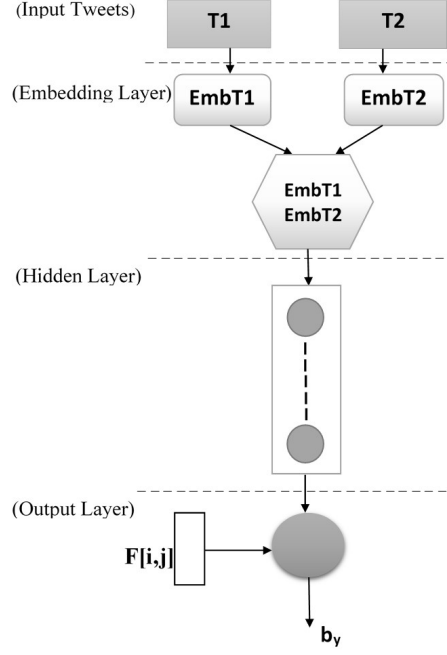


Figure 5.1: Pairwise learning model: network architecture.

Table 5.1: Parameters of the pairwise learning model.

Parameter	Value
Optimizer	Adam [27]
Learning Rate	0.005
Loss function	Binary cross entropy

5.2.2 Deep Learning H2O

Experiments were conducted with DL H2O, since deep learning models and similar network architectures are commonly used in previous works [7, 28, 30, 47, 50, 58]. Data Set 1 was used as input; at first, the classifier was trained with the original

(prior to SMOTE training set) and tested with the original test set, and then it was trained with the SMOTE over-sampled training set and tested with the original test set, as described in Section 4, subsection 4.1. Conducting experiments prior to SMOTE and after SMOTE was decided, in order to examine whether this technique improves the obtained results and the performance of the model. Additionally, these experiments were conducted with Feature Set 1, Feature Set 2 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance.

The following parameters were defined for all of the experiments: a. hyperbolic tangent (Tahn) function, was defined as the activation function, b. 2 hidden layers with 50 neurons each were used, c. epochs were set to 20; data set iterations, d. train samples for iteration were set to 16, e. adaptive learning rate algorithm was used; automatic combination of learning rate annealing and momentum training to avoid slow convergence, f. rho and epsilon parameters were set to $1.0E-8$ and 0.99, respectively, g. automatic standardization of data was enabled, h. L1 was set to $1.0E-5$; regularization method constraining the absolute value of the weights and dropping some of them, in order to reduce complexity and avoid overfitting, i. L2 was set to 0; regularization method constraining the sum of the squared weights. j. cross entropy was defined as the loss function; it is used when the model output consists of independent hypotheses (classification), while it is recommended when the target values are class labels and the data are imbalanced, k. Bernoulli was defined as the distribution function; it is recommended for binary classification, and l. the missing values were replaced with the mean value. A detailed description of these parameters is provided in Section 3, subsection 3.2.

Chapter 6

Results

In this Section, both the results of the experiments with each of the baseline classifiers, and the results of the experiments with each of the deep learning models are presented and discussed extensively.

6.1 Baseline Classifiers

The results of the experiments with each of the baseline classifiers and Data set 1 are presented and discussed. The overall performance is high in fake news and deception detection.

6.1.1 Naive Bayes

Experiments with the Naive Bayes classifier and Data Set 1 were conducted, achieving high overall performance in fake news and deception detection. Results prior to SMOTE and after SMOTE over-sampling with each of the feature sets (Feature Set 1 and 3) are described extensively. Conducting experiments prior to SMOTE and after SMOTE was decided, in order to examine whether this technique improves the obtained results and the performance of the model. Additionally, these experiments were conducted with Feature Set 1 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance.

Feature Set 1 (prior to SMOTE and after SMOTE)

The produced model achieved 100% Accuracy prior to SMOTE and 99.79% Accuracy after SMOTE over-sampling, which is high, compared to related work (Section 2).

Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.1 and Table 6.2.

Table 6.1: Naive Bayes Confusion Matrix. Feature Set 1. Prior to SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	0	100.00%
Predicted Fake	0	54	100.00%
Class Recall	100.00%	100.00%	

Table 6.2: Naive Bayes Confusion Matrix. Feature Set 1. After SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	1	99.76%
Predicted Fake	0	53	100.00%
Class Recall	100.00%	98.15%	

Several observations are made regarding the results and performance of the Naive Bayes model. Prior to SMOTE over-sampling, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 100%, and b. all examples labelled as “fake” are classified correctly, with a Class Recall of 100% and a Class Precision of 100%.

After SMOTE over-sampling, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 99.76%, and b. all examples labelled as “fake” are classified correctly except 1, with a Class Recall of 98.15% and a Class Precision of 100%. The wrong prediction concerning the only misclassified example occurs because, unlike the majority of the examples originally labelled as “fake”, it does not contain long sentences or big words (linguistic features), while containing 2 URLs (network feature).

Feature Set 3 (prior to SMOTE and after SMOTE)

The produced model achieved 100% Accuracy, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.3.

Several observations are made regarding the results and performance of the Naive Bayes model. Prior to SMOTE and after SMOTE over-sampling, the results are the same. More specifically, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 100%,

Table 6.3: Naive Bayes Confusion Matrix. Feature Set 3. Prior to and after SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	0	100.00%
Predicted Fake	0	54	100.00%
Class Recall	100.00%	100.00%	

and b. all examples labelled as “fake” are classified correctly, with a Class Recall of 100% and a Class Precision of 100%. No wrong predictions occur in this setting.

6.1.2 Random Forest

Experiments with the Random Forest classifier and Data Set 1 were conducted, achieving high overall performance in fake news and deception detection. Results prior to SMOTE and after SMOTE over-sampling with each of the feature sets (Feature Set 1 and 3) are described extensively. Conducting experiments prior to SMOTE and after SMOTE was decided, in order to examine whether this technique improves the obtained results and the performance of the model. Additionally, these experiments were conducted with Feature Set 1 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance.

Feature Set 1 (prior to SMOTE and after SMOTE)

The produced model achieved 98.52% Accuracy prior to SMOTE and 99.37% Accuracy after SMOTE over-sampling, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.4 and Table 6.5.

Table 6.4: Random Forest Confusion Matrix. Feature Set 1. Prior to SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	7	98.36%
Predicted Fake	0	47	100.00%
Class Recall	100.00%	87.04%	

Several observations are made regarding the results and performance of the Random Forest model. Prior to SMOTE over-sampling, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 98.36%, b. all examples labelled as “fake” are classified correctly except 7, with a Class Recall of 87.04% and a Class Precision of 100%.

Table 6.5: Random Forest Confusion Matrix. Feature Set 1. After SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	3	99.29%
Predicted Fake	0	51	100.00%
Class Recall	100.00%	94.44%	

After SMOTE over-sampling, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 99.29%, b. all examples labelled as “fake” are classified correctly except 3, with a Class Recall of 94.44% and a Class Precision of 100%. The wrong predictions concerning the 3 misclassified examples occur because, unlike the majority of the examples originally labelled as “fake”, they contain more than 7 words (linguistic feature) and 2 URLs (network feature).

Feature Set 3 (prior to SMOTE and after SMOTE)

The produced model achieved 98.73% Accuracy prior to SMOTE and 99.37% Accuracy after SMOTE over-sampling, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.6 and Table 6.7.

Table 6.6: Random Forest Confusion Matrix. Feature Set 3. Prior to SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	6	98.59%
Predicted Fake	0	48	100.00%
Class Recall	100.00%	88.89%	

Table 6.7: Random Forest Confusion Matrix. Feature Set 3. After SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	3	99.29%
Predicted Fake	0	51	100.00%
Class Recall	100.00%	94.44%	

Several observations are made regarding the results and performance of the Random Forest model. Prior to SMOTE over-sampling, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 98.59%, b. all examples labelled as “fake” are classified correctly except 3, with a Class Recall of 88.89% and a Class Precision of 100%.

After SMOTE over-sampling, the following are observed: a. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 99.29%, b. all examples labelled as “fake” are classified correctly except 3, with a Class Recall of 94.44% and a Class Precision of 100%. It is observed that this setting’s results and performance is exactly the same with the setting of the previous subsection, with Feature Set 1 after SMOTE over-sampling.

6.1.3 Generalized Linear Model H2O

Experiments with the GLM H2O classifier and Data Set 1 were conducted, achieving high overall performance in fake news and deception detection. Results prior to SMOTE and after SMOTE over-sampling with each of the feature sets (Feature Set 1 and 3) are described extensively. Conducting experiments prior to SMOTE and after SMOTE was decided, in order to examine whether this technique improves the obtained results and the performance of the model. Additionally, these experiments were conducted with Feature Set 1 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance.

Feature Set 1 (prior to SMOTE and after SMOTE)

The produced model achieved 99.37% Accuracy prior to SMOTE and 96.62% Accuracy after SMOTE over-sampling, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.8 and Table 6.9.

Table 6.8: GLM H2O Confusion Matrix. Feature Set 1. Prior to SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	3	99.29%
Predicted Fake	0	51	100.00%
Class Recall	100.00%	94.44%	

Table 6.9: GLM H2O Confusion Matrix. Feature Set 1. After SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	403	0	100.00%
Predicted Fake	16	54	77.14%
Class Recall	96.18%	100.00%	

Several observations are made regarding the results and performance of the GLM H2O model. Prior to SMOTE over-sampling, the following are observed: a. all examples labelled as “fake” are classified correctly except 3, with a Class Recall of 94.44% and a Class Precision of 100%, b. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 99.29%.

After SMOTE over-sampling, the following are observed: a. all examples labelled as “fake” are classified correctly, with a Class Recall of 100% and a Class Precision of 77.14%, b. all examples labelled as “real” are classified correctly except 16, with a Class Recall of 96.18% and a Class Precision of 100%. The wrong predictions concerning the 16 misclassified examples occur because, unlike the majority of the examples originally labelled as “real”, they contain less than 10 words and has a Flesh-Kincaid level less than 20 (linguistic features), while containing less than 2 URLs and have 0 likes (network features).

Feature Set 3 (prior to SMOTE and after SMOTE)

The produced model achieved 99.37% Accuracy prior to SMOTE and 96.62% Accuracy after SMOTE over-sampling, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.10 and Table 6.11.

Table 6.10: GLM H2O Confusion Matrix. Feature Set 3. Prior to SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	3	99.29%
Predicted Fake	0	51	100%
Class Recall	100.00%	94.44%	

Table 6.11: GLM H2O Confusion Matrix. Feature Set 3. After SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	403	0	100.00%
Predicted Fake	16	54	77.14%
Class Recall	96.18%	100.00%	

Several observations are made regarding the results and performance of the GLM H2O model. Prior to SMOTE over-sampling, the following are observed: a. all examples labelled as “fake” are classified correctly except 3, with a Class Recall of 94.44% and a Class Precision of 100%, b. all examples labelled as “real” are classified

correctly, with a Class Recall of 100% and a Class Precision of 99.29%. It is observed that this setting's results and performance is exactly the same with the setting of the previous subsection, with Feature Set 1 prior to SMOTE over-sampling.

After SMOTE over-sampling, the following are observed: a. all examples labelled as “fake” are classified correctly, with a Class Recall of 100% and a Class Precision of 77.14%, b. all examples labelled as “real” are classified correctly except 16, with a Class Recall of 96.18% and a Class Precision of 100%. It is observed that this setting's results and performance is exactly the same with the setting of the previous subsection, with Feature Set 1 after SMOTE over-sampling.

6.2 Deep Learning Models

The results of the experiments with each of the deep learning models and Data set 1 (DL H2O) and Data Set 2 (pairwise learning models) are presented and discussed. The overall performance is high in fake news and deception detection.

6.2.1 Pairwise Learning

Results of the experiments with the pairwise learning models and Data Set 2 and Feature Set 2 are shown in Table 6.12. For the first set of experiments ($T1$: real header and $T2$: real text, fake text, fake header), the following are observed: prior to SMOTE over-sampling, Precision is 97% for real tweets and 100% for fake tweets, and Recall is 95% for real tweets and 74% for fake tweets, while after SMOTE over-sampling, Precision is 100% for real tweets and 100% for fake tweets, and Recall is 100% for real tweets and 96% for fake tweets. For the second set of experiments ($T1$: real text and $T2$: real header, fake text, fake header), the following are observed: prior to SMOTE over-sampling, Precision is 99% for real tweets and 100% for fake tweets and Recall is 97% for real tweets and 93% for fake tweets, while after SMOTE over-sampling, Precision is 100% for real tweets and 100% for fake tweets and Recall is 96% for real tweets and 96% for fake tweets.

Several observations are made regarding the results and performance of the pairwise learning models. For both sets of experiments, the performance is increased after SMOTE over-sampling. The best performance is achieved for batch size 16 and 20 epochs. More specifically, for the second set of experiments, the performance is better than that of the first set of experiments prior to SMOTE over-sampling. Con-

Table 6.12: Performance for pairwise learning experiments. Prior to and after SMOTE.

Set 1			Set 2	
Tweet	Real	Fake	Real	Fake
Prior to SMOTE 2,363 tweets				
Precision	97%	100%	99%	100%
Recall	95%	74%	97%	93%
Total Accuracy	95%		94%	
Average F1 score	99%		98%	
After SMOTE 3,766 tweets				
Precision	100%	100%	100%	100%
Recall	100%	96%	96%	96%
Total Accuracy	98%		97%	
Average F1 score	100%		99%	

sequently, the setting of the second set of experiments is slightly more efficient for fake news detection, as it does not require SMOTE over-sampling to obtain higher performance results. The results of the second set of experiments also show that the real text ($T1$) is highly correlated with data ($T2$), compared to the respective correlation of the real header ($T1$) with data ($T2$) of the first set of experiments. The latter correlation is slightly increased after SMOTE over-sampling. As a result, the number of data may affect the performance of the framework of the first set of experiments, while the framework of the second set of experiments is more efficient in detecting fake news tweets (minority class).

6.2.2 Deep Learning H2O

Experiments with the DL H2O classifier and Data Set 1 were conducted, achieving high overall performance in fake news and deception detection. Results prior to SMOTE and after SMOTE over-sampling with each of the feature sets (Feature Set 1, 2 and 3) are described extensively. Conducting experiments prior to SMOTE and after SMOTE was decided, in order to examine whether this technique improves the obtained results and the performance of the model. Additionally, these experiments were conducted with Feature Set 1, Feature Set 2 and Feature Set 3, in order to examine the impact of different features to the obtained results and achieved performance.

Feature Set 1 (prior to SMOTE and after SMOTE)

The produced model achieved 99.79% Accuracy, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.13.

Table 6.13: DL H2O Confusion Matrix. Feature Set 1. Prior to and after SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	1	99.76%
Predicted Fake	0	53	100.00%
Class Recall	100.00%	98.15%	

Several observations are made regarding the results and performance of the DL H2O model. Prior to SMOTE and after SMOTE over-sampling, the results are the same. More specifically, the following are observed: a. all examples labelled as “fake” are classified correctly except 1, with a Class Recall of 98.15% and a Class Precision of 100%, b. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 99.76%. Other metrics are MSE: 4.156979E-5, RMSE: 0.0064474638 and R^2 : 0.99959207, prior to SMOTE, and MSE: 9.7356955E-5, RMSE: 0.009866963 and R^2 : 0.99961054, after SMOTE.

Feature Set 2 (prior to SMOTE and after SMOTE)

The produced model achieved 99.58% Accuracy prior to SMOTE and 100% Accuracy after SMOTE over-sampling, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.14 and Table 6.15.

Table 6.14: DL H2O Confusion Matrix. Feature Set 2. Prior to SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	2	99.52%
Predicted Fake	0	52	100.00%
Class Recall	100.00%	96.30%	

Several observations are made regarding the results and performance of the DL H2O model. Prior to SMOTE over-sampling, the following are observed: a. all examples labelled as “fake” are classified correctly except 2, with a Class Recall of 96.30% and a Class Precision of 100%, b. all examples labelled as “real” are classified

Table 6.15: DL H2O Confusion Matrix. Feature Set 2. After SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	0	100.00%
Predicted Fake	0	54	100.00%
Class Recall	100.00%	100.00%	

correctly, with a Class Recall of 100% and a Class Precision of 99.52%. Other metrics are MSE: 7.737548E-7, RMSE: 8.7963336E-4 and R^2 : 0.99999243.

After SMOTE over-sampling, the following are observed: a. all examples labelled as “fake” are classified correctly, with a Class Recall of 100% and a Class Precision of 100%, b. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 100%. Other metrics are MSE: 1.2015057E-5, RMSE: 0.0034662741 and R^2 : 0.99995196.

Feature Set 3 (prior to SMOTE and after SMOTE)

The produced model achieved 99.79% Accuracy, which is high, compared to related work (Section 2). Precision and Recall for each label are also high both prior to SMOTE and after SMOTE over-sampling, as shown in Table 6.16.

Table 6.16: DL H2O Confusion Matrix. Feature Set 3. Prior to and after SMOTE.

	True Real	True Fake	Class Precision
Predicted Real	419	1	99.76%
Predicted Fake	0	53	100.00%
Class Recall	100.00%	98.15%	

Several observations are made regarding the results and performance of the DL H2O model. Prior to SMOTE and after over-sampling, the results are the same. More specifically, the following are observed: a. all examples labelled as “fake” are classified correctly except 1, with a Class Recall of 98.15% and a Class Precision of 100%, b. all examples labelled as “real” are classified correctly, with a Class Recall of 100% and a Class Precision of 99.76%. Other metrics are MSE: 1.0034228E-4, RMSE: 0.0100171 and R^2 : 0.9990153, prior to SMOTE, and MSE: 1.9147113E-4, RMSE: 0.01383731 and R^2 : 0.99923414, after SMOTE.

Chapter 7

Conclusions

In this Section, the present thesis is concluded and the main findings are presented. The results are discussed extensively. The research questions, as presented in Section 1, are answered, since the results are correlated with them. The contribution is described in detail in subsection 7.1. Finally, certain directions for future work are provided in subsection 7.2.

7.1 Contribution

The following research questions have emerged:

- **Research Question 1:** Is it possible to design a methodology for creating and validating a real and fake news data set, consisting of social web text?
- **Research Question 2:** Which features can be extracted from such a data set? Which (type of) features are more suitable and determinant for efficient fake news and deception detection?
- **Research Question 3:** How is the performance of commonly used machine learning and NLP models for fake news and deception detection affected when the novel data set is used as input?
- **Research Question 4:** How is the performance of pairwise learning models created for different tasks (e.g. machine translation) affected when applied for fake news and deception detection, along with the novel data set as input? What kind of input can be used for creating an efficient pairwise setting?

Significant contribution regarding these research questions is offered in the present thesis, as well as in earlier work conducted by the author [39, 40], published during her Master’s studies. In the present thesis, the approaches regarding data, features, methodology, experiments, results and conclusions of the earlier works conducted by the author [39, 40] are gathered and presented as a more complete framework for fake news and deception detection. Contribution and results regarding the research questions is discussed in detail in the following paragraphs.

In earlier work conducted by the author [40], a novel, original, publicly available data set is created [21]. It consists of 2,366 tweets in English, regarding the events in Hong Kong protests of August, 2019. Processing of fake news tweets provided by Twitter, and collecting and validating real news tweets for these events are analyzed thoroughly, proposing a best practice setting for developing fake and real news data sets. Both fake news and real news are represented in this data set. Data imbalance is addressed successfully by using smoothing parameters and techniques, such as SMOTE over-sampling.

Multimodal input is used for the machine learning experiments, containing string-based and morphological-textual features (number of syllables, number of long sentences etc.) and higher-level linguistic-oriented features (Flesh-Kincaid level, adverbs-adjectives rate etc.) extracted from the unstructured tweet text, and network-oriented features (Feature Set 1). Additionally, several features are identified as more suitable and determinant for fake news and deception detection from tweets. These features, sorted from most to least determinant, are the following network features: a. user id, b. account creation date, c. following count, d. user display name, and e. user screen name. In general, experiments with feature sets using both linguistic-oriented and network-oriented features obtain the best results, regardless of the machine learning model. Consequently, multimodal input is considered suitable for fake news and deception detection tasks. Finally, the performance and obtained results of baseline classifiers are evaluated. These models are commonly used in related work and their performance with the novel data set as input is examined.

In earlier work conducted by the author [39], tweet vectors are used as input for the pairwise learning models for the first time. Every pair consists of one vector with real tweet data and one vector with real or fake tweet data. The correct classification of the second vector depends on the similarity or diversity when compared to the first vector.

Pairwise learning models are designed in an innovative way, in order to be used for

the first time for fake news and deception detection. It combines and fuses multimodal input at various layers, enabling flexibility and high performance. The input may be either news text or news header. Other researchers have used news articles headers and text as the pair of inputs. However, there is no precedent that categorizes tweets to headers and text based on their linguistic structure. This categorization is outlined in this thesis, in order to compare the pairwise learning models' performance with various input types. An additional feature set (Feature Set 2) is created, based on the feature set (Feature Set 1) and feature selection performed in earlier work conducted by the author [40].

The results of the pairwise learning models for fake news and deception detection are promising, and overall performance is high. For the first set of experiments ($T1$: real header and $T2$: real text, fake text, fake header), Accuracy (values ranging from 95% to 98%), Precision (values ranging from 97% to 100%), Recall (values ranging from 74% to 100%) and F1-score (values ranging from 99% to 100%) are increased after SMOTE over-sampling. For the second set of experiments ($T1$: real text and $T2$: real header, fake text, fake header), Accuracy (values ranging from 94% to 97%), Precision (values ranging from 99% to 100%), Recall (values ranging from 93% to 97%) and F1-score (values ranging from 98% to 99%) are increased after SMOTE over-sampling.

For the second set of experiments, the performance is better than that of the first set of experiments prior to SMOTE over-sampling. Consequently, the setting of the second set of experiments is slightly more efficient for fake news detection, as it does not require SMOTE over-sampling to obtain higher performance results. The results of the second set of experiments also show that the real text ($T1$) is highly correlated with data ($T2$), compared to the respective correlation of the real header ($T1$) with data ($T2$) of the first set of experiments. The latter correlation is slightly increased after SMOTE over-sampling. As a result, the number of data may affect the performance of the framework of the first set of experiments, while the framework of the second set of experiments is more efficient in detecting fake news tweets (minority class).

In the present thesis, an extensive literature review, including the most recent works regarding fake news and deception detection with machine learning and NLP methods is provided. Previous works with different types of data (news articles, tweets, etc.), various categories of features (network, linguistic, etc.), as well as the most efficient network architectures and machine learning models are described in

detail. Literature review provides a better and deeper understanding of the research field, while offering a benchmark regarding classification performance in fake news and deception detection.

The performance and obtained results of the machine learning models of the earlier work conducted by the author [40], namely Naive Bayes and Random Forest, are evaluated with the data set (Data Set 1, Feature Set 1) prior to SMOTE over-sampling as input. These models are commonly used in related work. Their performance with the data set prior to SMOTE as input is examined and compared to their performance of after SMOTE over-sampling. Prior to SMOTE over-sampling, baseline classifiers managed to predict the fake news and real news with high Accuracy, compared with previous works. More specifically, Naive Bayes achieves 100% Accuracy, and Random Forest achieves 98.52% Accuracy. Precision and Recall are also high, from 98% to 100%, and 87% to 100%, respectively. After SMOTE over-sampling, baseline classifiers managed to predict the fake news and real news with high Accuracy, compared with previous works. More specifically, Naive Bayes achieves 99.79% Accuracy, and Random Forest achieves 99.37% Accuracy. Precision and Recall are also high, from 99% to 100%, and 94% to 100%, respectively. It is observed that SMOTE over-sampling improves the performance of Random Forest, though not the performance of Naive Bayes, which does not classify correctly 1 fake news learning example.

An additional feature set (Feature Set 3) is created, based on the feature sets (Feature Set 1 and feature Set 2) and feature selection performed in both of the author's earlier works [39, 40]. Experiments with different feature sets are conducted and their effect on the machine learning models' performance is examined. Prior to SMOTE and after SMOTE over-sampling, Naive Bayes obtains the same results with Feature Set 3. More specifically, 100% Accuracy is achieved, while Precision and Recall are also high (both 100%). No wrong predictions occur on this setting. Regarding Random Forest and Feature Set 3, prior to SMOTE, 98.73% Accuracy is achieved. Precision and Recall are also high, from 98% to 100%, and 88% to 100%, respectively. After SMOTE over-sampling, evaluation metrics are improved. More specifically, 99.37% Accuracy is achieved, while Precision and Recall range from 99% to 100% and 94% to 100%, respectively.

The performance and obtained results of more machine learning models, namely GLM H2O (baseline classifier) and DL H2O (deep learning) are evaluated. Their performance with the novel data set as input is examined and compared to the performance of the models in earlier works conducted by the author [39, 40].

With Feature Set 1, GLM H2O achieves 99.37% Accuracy prior to SMOTE and 96.62% Accuracy after SMOTE over-sampling. Precision and Recall are also high, both prior to SMOTE (from 99% to 100%, and 94% to 100%, respectively) and after SMOTE over-sampling (from 77% to 100%, and 96% to 100%, respectively). With Feature Set 3, GLM H2O obtains exactly the same Accuracy and results. It is observed that the performance is worsened after SMOTE over-sampling. However, the only misclassified learning examples are real news, and all of the fake news learning examples are classified correctly, in contrast with prior to SMOTE over-sampling, where all real news learning examples are classified correctly, and 3 fake news learning examples are misclassified. In conclusion, despite performance after SMOTE over-sampling being worse, the model is actually more efficient in classifying fake news learning examples correctly (it is better at detecting fake news).

DL H2O achieves 99.79% Accuracy with Feature Set 1, regardless of whether SMOTE over-sampling was applied, though MSE, RMSE and R^2 are decreased after SMOTE over-sampling. With Feature Set 2, it achieves 99.58% Accuracy prior to SMOTE and 100% Accuracy after SMOTE over-sampling. With Feature Set 3, it achieves 99.79% Accuracy, regardless of whether SMOTE over-sampling was applied, though MSE, RMSE and R^2 are decreased after SMOTE over-sampling. Precision and Recall are also high, from 99% to 100%, and 96% to 100%, respectively.

The performance of all machine learning models, with every Feature Set, prior to SMOTE and after SMOTE over-sampling, is presented in Table 7.1 (Feature Set 1), Table 7.2 (Feature Set 2) and Table 7.3 (Feature Set 3).

Table 7.1: Performance of all models. Feature Set 1. Prior to SMOTE (PS) and after SMOTE (AS). NB: Naive Bayes, RF: Random Forest, GLM: GLM H2O, DL: DL H2O, Acc: Accuracy, Prec: Precision, Rec: Recall, R: Real news, F: Fake news.

	PS					AS				
	Acc	Prec		Rec		Acc	Prec		Rec	
		R	F	R	F		R	F	R	F
NB	100%	100%	100%	100%	100%	99.79%	99.76%	100%	100%	98.15%
RF	98.52%	98.36%	100%	100%	87.04%	99.37%	99.29%	100%	100%	94.44%
GLM	99.37%	99.29%	100%	100%	94.44%	96.62%	100%	77.14%	96.18%	100%
DL	99.79%	99.76%	100%	100%	98.15%	99.79%	99.76%	100%	100%	98.15%

Table 7.2: Performance of all models. Feature Set 2. Prior to SMOTE (PS) and after SMOTE (AS). PW-1: pairwise learning model (Set 1), PW-2: pairwise learning model (Set 2), DL: DL H2O, Acc: Accuracy, Prec: Precision, Rec: Recall, R: Real news, F: Fake news.

	PS					AS				
	Acc	Prec		Rec		Acc	Prec		Rec	
		R	F	R	F		R	F	R	F
PW-1	95%	97%	100%	95%	74%	98%	100%	100%	100%	96%
PW-2	94%	99%	100%	97%	93%	97%	100%	100%	96%	96%
DL	99.58%	99.52%	100%	100%	96.30%	100%	100%	100%	100%	100%

Table 7.3: Performance of all models. Feature Set 3. Prior to SMOTE (PS) and after SMOTE (AS). NB: Naive Bayes, RF: Random Forest, GLM: GLM H2O, DL: DL H2O, Acc: Accuracy, Prec: Precision, Rec: Recall, R: Real news, F: Fake news.

	PS					AS				
	Acc	Prec		Rec		Acc	Prec		Rec	
		R	F	R	F		R	F	R	F
NB	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
RF	98.73%	98.59%	100%	100%	88.89%	99.37%	99.29%	100%	100%	94.44%
GLM	99.37%	99.29%	100%	100%	94.44%	96.62%	100%	77.14%	96.18%	100%
DL	99.79%	99.76%	100%	100%	98.15%	99.79%	99.76%	100%	100%	98.15%

7.2 Future Work

Automated fake news and deception detection poses a number of challenges for machine learning, text mining and NLP research fields. The present thesis outlines several directions for future work, as described below:

- Experiments with the proposed machine learning models could be conducted with more data, or data from different social networks and media, such as Facebook, blogs and wikis;
- The performance of the proposed machine learning models could be examined with data that contain other types of fake news, such as spam messages;
- The performance of the proposed machine learning models could be examined with other types of data, such as multimedia content, photos and videos;
- The performance of the proposed machine learning models could be examined with other types of features and different feature sets;

- Different pairwise learning models configurations, such as different kinds of neural network layers and pre-trained embeddings, could be tested to improve the obtained results and overall performance for fake news and deception detection;
- Different pairs of inputs could be used for pairwise learning models, to examine different aspects of tweet text;
- Finally, the features that were defined as more suitable and determinant for fake news and deception detection could be further examined, in order to identify what makes them so.

Appendix A

Abbreviations

CNN	Convolutional Neural Network
DL H2O	Deep Learning H2O
GLM H2O	Generalized Linear Model H2O
KNN	k-Nearest Neighbor
LIWC	Linguistic Inquiry and Word Count
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POS	Part-Of-Speech
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency

Appendix B

XML: Baseline classifiers experiments (RapidMiner Studio)

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="246" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data_management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING<br/></description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="447" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
      </operator>
    </process>
  </operator>
</process>
```

```

<parameter key="time_zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data_set_meta_data_information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double_array"/>
<parameter key="data_management" value="auto"/>
<description align="center" color="transparent" colored="false" width="126">TEST</description>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="66" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
  <parameter key="repository_entry" value="../../../../data/TRAINING"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="naive_bayes" compatibility="9.8.000" expanded="true" height="82" name="Naive Bayes" width="90" x="313" y="34">
  <parameter key="laplace_correction" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="66" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Naive Bayes" to_port="training set"/>
<connect from_op="Naive Bayes" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink result 1" spacing="0"/>
<portSpacing port="sink result 2" spacing="0"/>
<portSpacing port="sink result 3" spacing="0"/>
</process>
</operator>
/process>

```

Figure B.1: Experiments with Naive Bayes (Feature Set 1, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
  </context>
  <macros/>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="246" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double array"/>
        <parameter key="data_management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING<br/></description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="447" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
      </operator>
    </process>
  </operator>
  <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
    <parameter key="repository_entry" value="../../data/SMOTE TRAINING"/>
  </operator>
  <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
    <parameter key="attribute_filter_type" value="subset"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Fleish-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Urls|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="attribute value"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="time"/>
    <parameter key="block_type" value="attribute block"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value matrix row start"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
  </operator>
  <operator activated="true" class="naive_bayes" compatibility="9.8.000" expanded="true" height="82" name="Naive Bayes" width="90" x="313" y="34">
    <parameter key="laplace_correction" value="true"/>
  </operator>
  <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
    <parameter key="repository_entry" value="../../data/TEST"/>
  </operator>

```

```

<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179"
y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words
|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_exopt_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Naive Bayes" to_port="training set"/>
<connect from_op="Naive Bayes" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink result 1" spacing="0"/>
<portSpacing port="sink result 2" spacing="0"/>
<portSpacing port="sink result 3" spacing="0"/>
</process>
</operator>

```

Figure B.2: Experiments with Naive Bayes (Feature Set 1, after SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process duration for mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="246" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data set meta data information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING<br/></description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="447" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
      </operator>
    </process>
  </operator>

```

```

<list key="annotations"/>
<parameter key="date_format" value=""/>
<parameter key="time_zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data_set_meta_data_information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double_array"/>
<parameter key="data_management" value="auto"/>
<description align="Center" color="transparent" colored="false" width="126">TEST</description>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
  <parameter key="repository_entry" value="../../../../data/TRAINING"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="naive_bayes" compatibility="9.8.000" expanded="true" height="82" name="Naive Bayes" width="90" x="313" y="34">
  <parameter key="laplace_correction" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>

<operator activated="false" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="112" y="391">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (3)" width="90" x="179" y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>

```

Figure B.3: Experiments with Naive Bayes (Feature Set 3, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="246" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data_management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING<br/></description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="447" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>

```

```

    <parameter key="date_format" value=""/>
    <parameter key="time_zone" value="SYSTEM"/>
    <parameter key="locale" value="English (United States)"/>
    <parameter key="read_all_values_as_polynomial" value="false"/>
    <list key="data_set_meta_data_information"/>
    <parameter key="read_not_matching_values_as_missings" value="true"/>
    <parameter key="datamanagement" value="double_array"/>
    <parameter key="data_management" value="auto"/>
    <description align="center" color="transparent" colored="false" width="126">TEST</description>
  </operator>
  <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE_TRAINING" width="90" x="45" y="34">
    <parameter key="repository_entry" value="../../../../data/SMOTE_TRAINING"/>
  </operator>
  <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
    <parameter key="attribute_filter_type" value="subset"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Fleish-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Urls|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="attribute_value"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="time"/>
    <parameter key="block_type" value="attribute_block"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_matrix_row_start"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
  </operator>
  <operator activated="true" class="naive_bayes" compatibility="9.8.000" expanded="true" height="82" name="Naive Bayes" width="90" x="313" y="34">
    <parameter key="laplace_correction" value="true"/>
  </operator>
  <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
    <parameter key="repository_entry" value="../../../../data/TEST"/>
  </operator>

```

```

<operator activated="false" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="112"
y="391">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words
|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Urls|Num_words|Rate_Adv_adj|Retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (3)" width="90" x="179"
y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences
|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Urls|Num_words|Rate_Adv_adj|Retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>

```

Figure B.4: Experiments with Naive Bayes (Feature Set 3, after SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="380" y="391">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING</description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="179" y="391">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
      </operator>
    </process>
  </operator>

```

```

<list key="annotations"/>
<parameter key="date_format" value=""/>
<parameter key="time_zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data_set_meta_data_information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double array"/>
<parameter key="data_management" value="auto"/>
<description align="center" color="transparent" colored="false" width="126">TEST</description>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
  <parameter key="repository_entry" value="../../../../data/TRAINING"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="true"/>
</operator>
<operator activated="true" class="concurrency:parallel_random_forest" compatibility="9.8.000" expanded="true" height="103" name="Random Forest" width="90" x="313" y="34">
  <parameter key="number_of_trees" value="100"/>
  <parameter key="criterion" value="gain_ratio"/>
  <parameter key="maximal_depth" value="50"/>
  <parameter key="apply_pruning" value="false"/>
  <parameter key="confidence" value="0.1"/>

  <parameter key="apply_prepruning" value="false"/>
  <parameter key="minimal_gain" value="0.01"/>
  <parameter key="minimal_leaf_size" value="2"/>
  <parameter key="minimal_size_for_split" value="4"/>
  <parameter key="number_of_prepruning_alternatives" value="3"/>
  <parameter key="random_splits" value="false"/>
  <parameter key="guess_subset_ratio" value="false"/>
  <parameter key="subset_ratio" value="0.2"/>
  <parameter key="voting_strategy" value="majority vote"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="enable_parallel_execution" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="246" y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="true"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>

  <operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
    <parameter key="use_example_weights" value="true"/>
  </operator>
  <connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
  <connect from_op="Select Attributes" from_port="example set output" to_op="Random Forest" to_port="training set"/>
  <connect from_op="Random Forest" from_port="model" to_op="Apply Model" to_port="model"/>
  <connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
  <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
  <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
  <connect from_op="Apply Model" from_port="model" to_port="result 3"/>
  <connect from_op="Performance" from_port="performance" to_port="result 1"/>
  <connect from_op="Performance" from_port="example set" to_port="result 2"/>
  <portSpacing port="source input 1" spacing="0"/>
  <portSpacing port="sink result 1" spacing="0"/>
  <portSpacing port="sink result 2" spacing="0"/>
  <portSpacing port="sink result 3" spacing="0"/>
  <portSpacing port="sink result 4" spacing="0"/>
</process>
</operator>
</process>

```

Figure B.5: Experiments with Random Forest (Feature Set 1, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process duration for mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="380" y="391">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING</description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="179" y="391">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>

```

```

    <parameter key="date_format" value=""/>
    <parameter key="time_zone" value="SYSTEM"/>
    <parameter key="locale" value="English (United States)"/>
    <parameter key="read_all_values_as_polynomial" value="false"/>
    <list key="data_set_meta_data_information"/>
    <parameter key="read_not_matching_values_as_missings" value="true"/>
    <parameter key="datamanagement" value="double_array"/>
    <parameter key="data management" value="auto"/>
    <description align="center" color="transparent" colored="false" width="126">TEST</description>
  </operator>
  <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
    <parameter key="repository_entry" value=".../data/SMOTE TRAINING"/>
  </operator>
  <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
    <parameter key="attribute_filter_type" value="subset"/>
    <parameter key="attribute" value=""/>
    <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Fleish-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
    <parameter key="use_except_expression" value="false"/>
    <parameter key="value_type" value="attribute_value"/>
    <parameter key="use_value_type_exception" value="false"/>
    <parameter key="except_value_type" value="time"/>
    <parameter key="block_type" value="attribute_block"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_matrix_row_start"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="true"/>
  </operator>
  <operator activated="true" class="concurrency:parallel_random_forest" compatibility="9.8.000" expanded="true" height="103" name="Random Forest" width="90" x="313" y="34">
    <parameter key="number_of_trees" value="100"/>
    <parameter key="criterion" value="gain_ratio"/>
    <parameter key="maximal_depth" value="50"/>
    <parameter key="apply_pruning" value="false"/>
    <parameter key="confidence" value="0.1"/>
    <parameter key="apply_prepruning" value="false"/>

```

```

<parameter key="minimal_gain" value="0.01"/>
<parameter key="minimal_leaf_size" value="2"/>
<parameter key="minimal_size_for_split" value="4"/>
<parameter key="number_of_prepruning_alternatives" value="3"/>
<parameter key="random_splits" value="false"/>
<parameter key="guess_subset_ratio" value="false"/>
<parameter key="subset_ratio" value="0.2"/>
<parameter key="voting_strategy" value="majority vote"/>
<parameter key="use_local_random_seed" value="false"/>
<parameter key="local_random_seed" value="1992"/>
<parameter key="enable_parallel_execution" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value=".../data/TEST"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Urls|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value matrix row start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="true"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>

<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Random Forest" to_port="training set"/>
<connect from_op="Random Forest" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 3"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink result 1" spacing="0"/>
<portSpacing port="sink result 2" spacing="0"/>
<portSpacing port="sink result 3" spacing="0"/>
<portSpacing port="sink result 4" spacing="0"/>
</process>
</operator>
</process>

```

Figure B.6: Experiments with Random Forest (Feature Set 1, after SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="380" y="391">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data_management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING</description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="112" y="391">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
      </operator>
    </process>
  </operator>
</process>

```

```

<parameter key="date format" value=""/>
<parameter key="time zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data set meta data information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double array"/>
<parameter key="data_management" value="auto"/>
<description align="center" color="transparent" colored="false" width="126">TEST</description>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
  <parameter key="repository_entry" value="../../../../data/TRAINING"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="true"/>
</operator>
<operator activated="true" class="concurrency:parallel_random_forest" compatibility="9.8.000" expanded="true" height="103" name="Random Forest" width="90" x="313" y="34">
  <parameter key="number of trees" value="100"/>
  <parameter key="criterion" value="gain ratio"/>
  <parameter key="maximal_depth" value="50"/>
  <parameter key="apply_pruning" value="false"/>
  <parameter key="confidence" value="0.1"/>
  <parameter key="apply_prepruning" value="false"/>
  <parameter key="minimal_gain" value="0.01"/>
  <parameter key="minimal_leaf_size" value="2"/>
  <parameter key="minimal_size_for_split" value="4"/>
  <parameter key="number_of_prepruning_alternatives" value="3"/>
  <parameter key="random_splits" value="false"/>
  <parameter key="guess_subset_ratio" value="false"/>
  <parameter key="subset_ratio" value="0.2"/>
  <parameter key="voting_strategy" value="majority vote"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="enable_parallel_execution" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>

<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Random Forest" to_port="training set"/>
<connect from_op="Random Forest" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 3"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink result 1" spacing="0"/>
<portSpacing port="sink result 2" spacing="0"/>
<portSpacing port="sink result 3" spacing="0"/>
<portSpacing port="sink result 4" spacing="0"/>
</process>
</operator>
/process>

```

Figure B.7: Experiments with Random Forest (Feature Set 3, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
<context>
<input/>
<output/>
<macros/>
</context>
<operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
<parameter key="logverbosity" value="init"/>
<parameter key="random_seed" value="2001"/>
<parameter key="send_mail" value="never"/>
<parameter key="notification_email" value=""/>
<parameter key="process_duration_for_mail" value="30"/>
<parameter key="encoding" value="SYSTEM"/>
<process expanded="true">
<operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="380" y="391">
<parameter key="sheet_selection" value="sheet number"/>
<parameter key="sheet_number" value="1"/>
<parameter key="imported_cell_range" value="A1"/>
<parameter key="encoding" value="SYSTEM"/>
<parameter key="first_row_as_names" value="true"/>
<list key="annotations"/>
<parameter key="date_format" value=""/>
<parameter key="time_zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data_set_meta_data_information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double_array"/>
<parameter key="data_management" value="auto"/>
<description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING</description>
</operator>
<operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="179" y="391">
<parameter key="sheet_selection" value="sheet number"/>
<parameter key="sheet_number" value="1"/>
<parameter key="imported_cell_range" value="A1"/>
<parameter key="encoding" value="SYSTEM"/>
<parameter key="first_row_as_names" value="true"/>
<list key="annotations"/>
<parameter key="date_format" value=""/>
<parameter key="time_zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data_set_meta_data_information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double_array"/>
<parameter key="data_management" value="auto"/>
<description align="center" color="transparent" colored="false" width="126">TEST</description>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
<parameter key="repository_entry" value="../../../../data/SMOTE TRAINING"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
<parameter key="attribute filter type" value="subset"/>
<parameter key="attribute" value=""/>
<parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
<parameter key="use_except_expression" value="false"/>
<parameter key="value_type" value="attribute_value"/>
<parameter key="use_value_type_exception" value="false"/>
<parameter key="except_value_type" value="time"/>
<parameter key="block_type" value="attribute_block"/>
<parameter key="use_block_type_exception" value="false"/>
<parameter key="except_block_type" value="value_matrix_row_start"/>
<parameter key="invert_selection" value="false"/>
<parameter key="include_special_attributes" value="true"/>
</operator>
<operator activated="true" class="concurrency:parallel_random_forest" compatibility="9.8.000" expanded="true" height="103" name="Random Forest" width="90" x="313" y="34">
<parameter key="number_of_trees" value="100"/>
<parameter key="criterion" value="gain_ratio"/>
<parameter key="maximal_depth" value="50"/>
<parameter key="apply_pruning" value="false"/>
<parameter key="confidence" value="0.1"/>

```

```

<parameter key="apply_prepruning" value="false"/>
<parameter key="minimal_gain" value="0.01"/>
<parameter key="minimal_leaf_size" value="2"/>
<parameter key="minimal_size_for_split" value="4"/>
<parameter key="number_of_prepruning_alternatives" value="3"/>
<parameter key="random_splits" value="false"/>
<parameter key="guess_subset_ratio" value="false"/>
<parameter key="subset_ratio" value="0.2"/>
<parameter key="voting_strategy" value="majority vote"/>
<parameter key="use_local_random_seed" value="false"/>
<parameter key="local_random_seed" value="1992"/>
<parameter key="enable_parallel_execution" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value="../../data/TEST"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="239">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_exompt_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>

<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Random Forest" to_port="training set"/>
<connect from_op="Random Forest" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 3"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source_input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
<portSpacing port="sink_result 4" spacing="0"/>
</process>
</operator>
</process>

```

Figure B.8: Experiments with Random Forest (Feature Set 3, after SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
<context>
  <input/>
  <output/>
  <macros/>
</context>
<operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
  <parameter key="logverbosity" value="init"/>
  <parameter key="random_seed" value="2001"/>
  <parameter key="send_mail" value="never"/>
  <parameter key="notification_email" value=""/>
  <parameter key="process_duration_for_mail" value="30"/>
  <parameter key="encoding" value="SYSTEM"/>
  <process expanded="true">
    <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="246" y="442">
      <parameter key="sheet_selection" value="sheet number"/>
      <parameter key="sheet_number" value="1"/>
      <parameter key="imported_cell_range" value="A1"/>
      <parameter key="encoding" value="SYSTEM"/>
      <parameter key="first_row_as_names" value="true"/>
      <list key="annotations"/>
      <parameter key="date_format" value=""/>
      <parameter key="time_zone" value="SYSTEM"/>
      <parameter key="locale" value="English (United States)"/>
      <parameter key="read_all_values_as_polynomial" value="false"/>
      <list key="data set meta data information"/>
      <parameter key="read_not_matching values as missings" value="true"/>
      <parameter key="datamanagement" value="double_array"/>
      <parameter key="data management" value="auto"/>
      <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING<br/></description>
    </operator>
    <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="447" y="442">
      <parameter key="sheet_selection" value="sheet number"/>
      <parameter key="sheet_number" value="1"/>
      <parameter key="imported_cell_range" value="A1"/>
      <parameter key="encoding" value="SYSTEM"/>
      <parameter key="first_row_as_names" value="true"/>
      <list key="annotations"/>
    </operator>
  </process>
</operator>

```

```

<parameter key="date_format" value=""/>
<parameter key="time_zone" value="SYSTEM"/>
<parameter key="locale" value="English (United States)"/>
<parameter key="read_all_values_as_polynomial" value="false"/>
<list key="data_set_meta_data_information"/>
<parameter key="read_not_matching_values_as_missings" value="true"/>
<parameter key="datamanagement" value="double array"/>
<parameter key="data_management" value="auto"/>
<description align="center" color="transparent" colored="false" width="126">TEST</description>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
  <parameter key="repository_entry" value="../../../../data/TRAINING"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="h2o:generalized_linear_model" compatibility="9.8.000" expanded="true" height="124" name="Generalized Linear Model" width="90" x="313" y="34">
  <parameter key="family" value="AUTO"/>
  <parameter key="link" value="family_default"/>
  <parameter key="solver" value="AUTO"/>
  <parameter key="reproducible" value="false"/>
  <parameter key="maximum_number_of_threads" value="4"/>
  <parameter key="use_regularization" value="true"/>
  <parameter key="lambda_search" value="false"/>
  <parameter key="number_of_lambdas" value="0"/>
  <parameter key="lambda_min_ratio" value="0.0"/>
  <parameter key="early_stopping" value="true"/>
  <parameter key="stopping_rounds" value="3"/>
  <parameter key="stopping_tolerance" value="0.001"/>
  <parameter key="standardize" value="true"/>
  <parameter key="non-negative_coefficients" value="false"/>
  <parameter key="add_intercept" value="true"/>
  <parameter key="compute_p-values" value="false"/>
  <parameter key="remove_collinear_columns" value="false"/>
  <parameter key="missing_values_handling" value="MeanImputation"/>
  <parameter key="max_iterations" value="0"/>
  <parameter key="specify_beta_constraints" value="false"/>
  <list key="beta_constraints"/>
  <parameter key="max_runtime_seconds" value="0"/>
  <list key="expert_parameters"/>
</operator>
<operator activated="false" class="naive bayes" compatibility="9.8.000" expanded="true" height="82" name="Naive Bayes" width="90" x="648" y="340">
  <parameter key="laplace_correction" value="true"/>
</operator>
<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="238">
  <parameter key="attribute_filter_type" value="subset"/>
  <parameter key="attribute" value=""/>
  <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
  <parameter key="use_except_expression" value="false"/>
  <parameter key="value_type" value="attribute_value"/>
  <parameter key="use_value_type_exception" value="false"/>
  <parameter key="except_value_type" value="time"/>
  <parameter key="block_type" value="attribute_block"/>
  <parameter key="use_block_type_exception" value="false"/>
  <parameter key="except_block_type" value="value_matrix_row_start"/>
  <parameter key="invert_selection" value="false"/>
  <parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="238">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Generalized Linear Model" to_port="training set"/>
<connect from_op="Generalized Linear Model" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source_input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
</process>
</operator>
/process>

```

Figure B.9: Experiments with GLM H2O (Feature Set 1, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel" width="90" x="246" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data_management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">SMOTE_TRAINING<br/></description>
      </operator>
      <operator activated="false" class="read_excel" compatibility="9.8.000" expanded="true" height="68" name="Read Excel (2)" width="90" x="447" y="442">
        <parameter key="sheet_selection" value="sheet number"/>
        <parameter key="sheet_number" value="1"/>
        <parameter key="imported_cell_range" value="A1"/>
        <parameter key="encoding" value="SYSTEM"/>
        <parameter key="first_row_as_names" value="true"/>
        <list key="annotations"/>
        <parameter key="date_format" value=""/>
        <parameter key="time_zone" value="SYSTEM"/>
        <parameter key="locale" value="English (United States)"/>
        <parameter key="read_all_values_as_polynomial" value="false"/>
        <list key="data_set_meta_data_information"/>
        <parameter key="read_not_matching_values_as_missings" value="true"/>
        <parameter key="datamanagement" value="double_array"/>
        <parameter key="data_management" value="auto"/>
        <description align="center" color="transparent" colored="false" width="126">TEST</description>
      </operator>
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value=".../data/SMOTE_TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:generalized_linear_model" compatibility="9.8.000" expanded="true" height="124" name="Generalized Linear Model" width="90" x="313" y="34">
        <parameter key="family" value="AUTO"/>
        <parameter key="link" value="family default"/>
        <parameter key="solver" value="AUTO"/>
        <parameter key="reproducible" value="false"/>
        <parameter key="maximum_number_of_threads" value="4"/>
        <parameter key="use_regularization" value="true"/>
        <parameter key="lambda_search" value="false"/>
        <parameter key="number_of_lambdas" value="0"/>
        <parameter key="lambda_min_ratio" value="0.0"/>
        <parameter key="early_stopping" value="true"/>
        <parameter key="stopping_rounds" value="3"/>
        <parameter key="stopping_tolerance" value="0.001"/>
        <parameter key="standardize" value="true"/>
        <parameter key="non-negative_coefficients" value="false"/>
        <parameter key="add_intercept" value="true"/>
        <parameter key="compute_p-values" value="false"/>
        <parameter key="remove_collinear_columns" value="false"/>
        <parameter key="missing_values_handling" value="MeanImputation"/>
        <parameter key="max_iterations" value="0"/>
        <parameter key="specify_beta_constraints" value="false"/>
        <list key="beta_constraints"/>
        <parameter key="max_runtime_seconds" value="0"/>
        <list key="expert_parameters"/>
      </operator>
      <operator activated="false" class="naive_bayes" compatibility="9.8.000" expanded="true" height="82" name="Naive Bayes" width="90" x="648" y="340">
        <parameter key="laplace_correction" value="true"/>
      </operator>
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
        <parameter key="repository_entry" value=".../data/TEST"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="238">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute block"/>
      </operator>
    </process>
  </operator>

```

```

    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_matrix_row_start"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
  </operator>
  <operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
    <list key="application_parameters"/>
    <parameter key="create_view" value="false"/>
  </operator>
  <operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
    <parameter key="use_example_weights" value="true"/>
  </operator>
  <connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
  <connect from_op="Select Attributes" from_port="example set output" to_op="Generalized Linear Model" to_port="training set"/>
  <connect from_op="Generalized Linear Model" from_port="model" to_op="Apply Model" to_port="model"/>
  <connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
  <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
  <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
  <connect from_op="Performance" from_port="performance" to_port="result 1"/>
  <connect from_op="Performance" from_port="performance" to_port="example set" to_port="result 2"/>
  <portSpacing port="source input 1" spacing="0"/>
  <portSpacing port="sink result 1" spacing="0"/>
  <portSpacing port="sink result 2" spacing="0"/>
  <portSpacing port="sink result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure B.10: Experiments with GLM H2O (Feature Set 1, after SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random seed" value="2001"/>
    <parameter key="send mail" value="never"/>
    <parameter key="notification email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../../data/TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:generalized_linear_model" compatibility="9.8.000" expanded="true" height="124" name="Generalized Linear Model" width="90" x="313" y="34">
        <parameter key="family" value="AUTO"/>
      </operator>
      <parameter key="link" value="family default"/>
      <parameter key="solver" value="AUTO"/>
      <parameter key="reproducible" value="false"/>
      <parameter key="maximum number of threads" value="4"/>
      <parameter key="use regularization" value="true"/>
      <parameter key="lambda_search" value="false"/>
      <parameter key="number of lambdas" value="0"/>
      <parameter key="lambda_min_ratio" value="0.0"/>
      <parameter key="early_stopping" value="true"/>
      <parameter key="stopping_rounds" value="3"/>
      <parameter key="stopping_tolerance" value="0.001"/>
      <parameter key="standardize" value="true"/>
      <parameter key="non-negative coefficients" value="false"/>
      <parameter key="add_intercept" value="true"/>
      <parameter key="compute_p-values" value="false"/>
      <parameter key="remove collinear columns" value="false"/>
      <parameter key="missing values handling" value="MeanImputation"/>
      <parameter key="max iterations" value="0"/>
      <parameter key="specify beta constraints" value="false"/>
      <list key="beta constraints"/>
      <parameter key="max runtime seconds" value="0"/>
      <list key="expert parameters"/>
    </process>
    <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
      <parameter key="repository_entry" value="../../../../data/TEST"/>
    </operator>
    <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="246" y="238">
      <parameter key="attribute_filter_type" value="subset"/>
      <parameter key="attribute" value=""/>
      <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
      <parameter key="use_except_expression" value="false"/>
      <parameter key="value_type" value="attribute_value"/>
      <parameter key="use_value_type_exception" value="false"/>
      <parameter key="except_value_type" value="time"/>
    </operator>
  </process>

```

```

    <parameter key="block_type" value="attribute_block"/>
    <parameter key="use_block_type_exception" value="false"/>
    <parameter key="except_block_type" value="value_matrix_row_start"/>
    <parameter key="invert_selection" value="false"/>
    <parameter key="include_special_attributes" value="false"/>
  </operator>
  <operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
    <list key="application_parameters"/>
    <parameter key="create_view" value="false"/>
  </operator>
  <operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
    <parameter key="use_example_weights" value="true"/>
  </operator>
  <connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
  <connect from_op="Select Attributes" from_port="example set output" to_op="Generalized Linear Model" to_port="training set"/>
  <connect from_op="Generalized Linear Model" from_port="model" to_op="Apply Model" to_port="model"/>
  <connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
  <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
  <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
  <connect from_op="Performance" from_port="performance" to_port="result 1"/>
  <connect from_op="Performance" from_port="example set" to_port="result 2"/>
  <portSpacing port="source input 1" spacing="0"/>
  <portSpacing port="sink result 1" spacing="0"/>
  <portSpacing port="sink result 2" spacing="0"/>
  <portSpacing port="sink result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure B.11: Experiments with GLM H2O (Feature Set 3, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random seed" value="2001"/>
    <parameter key="send mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../../data/SMOTE TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:generalized_linear_model" compatibility="9.8.000" expanded="true" height="124" name="Generalized Linear Model" width="90" x="313" y="34">
        <parameter key="family" value="AUTO"/>
        <parameter key="link" value="family default"/>
        <parameter key="solver" value="AUTO"/>
        <parameter key="reproducible" value="false"/>
        <parameter key="maximum_number_of_threads" value="4"/>
        <parameter key="use_regularization" value="true"/>
        <parameter key="lambda_search" value="false"/>
        <parameter key="number_of_lambdas" value="0"/>
        <parameter key="lambda_min_ratio" value="0.0"/>
        <parameter key="early_stopping" value="true"/>
        <parameter key="stopping_rounds" value="3"/>
        <parameter key="stopping_tolerance" value="0.001"/>
        <parameter key="standardize" value="true"/>
        <parameter key="non-negative coefficients" value="false"/>
        <parameter key="add_intercept" value="true"/>
        <parameter key="compute_p-values" value="false"/>
        <parameter key="remove_collinear_columns" value="false"/>
        <parameter key="missing_values_handling" value="MeanImputation"/>
        <parameter key="max_iterations" value="0"/>
        <parameter key="specify_beta_constraints" value="false"/>
        <list key="beta_constraints"/>
        <parameter key="max_runtime_seconds" value="0"/>
        <list key="expert_parameters"/>
      </operator>
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="45" y="238">
        <parameter key="repository_entry" value="../../../../data/TEST"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes (2)" width="90" x="179" y="238">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
      </operator>
    </process>
  </operator>
</process>

```

```

<parameter key="block_type" value="attribute_block"/>
<parameter key="use_block_type_exception" value="false"/>
<parameter key="except_block_type" value="value_matrix_row_start"/>
<parameter key="invert_selection" value="false"/>
<parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="380" y="238">
  <list key="application parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="447" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Generalized Linear Model" to_port="training set"/>
<connect from_op="Generalized Linear Model" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Select Attributes (2)" to_port="example set input"/>
<connect from_op="Select Attributes (2)" from_port="example set output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<connect from_op="Performance" from_port="example set" to_port="result 2"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
</process>
</operator>
/process>

```

Figure B.12: Experiments with GLM H2O (Feature Set 3, after SMOTE).

Appendix C

XML: Deep learning models experiments (RapidMiner Studio)

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="false" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="112" y="340">
        <parameter key="repository_entry" value="../../../data/SMOTE TRAINING"/>
      </operator>
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="112" y="187">
        <parameter key="repository_entry" value="../../../data/TEST"/>
      </operator>
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../data/TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute block"/>
        <parameter key="use_block_type_exception" value="false"/>
      </operator>
    </process>
  </operator>
</process>
```

```

<parameter key="except_block_type" value="value_matrix_row_start"/>
<parameter key="invert_selection" value="false"/>
<parameter key="include_special_attributes" value="false"/>
</operator>
<operator activated="true" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning" width="90" x="313" y="34">
  <parameter key="activation" value="Tanh"/>
  <enumeration key="hidden_layer_sizes">
    <parameter key="hidden_layer_sizes" value="50"/>
    <parameter key="hidden_layer_sizes" value="50"/>
  </enumeration>
  <enumeration key="hidden_dropout_ratios">
  <parameter key="reproducible_(uses_1_thread)" value="false"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="epochs" value="20.0"/>
  <parameter key="compute_variable_importances" value="false"/>
  <parameter key="train_samples_per_iteration" value="16"/>
  <parameter key="adaptive_rate" value="true"/>
  <parameter key="epsilon" value="1.0E-8"/>
  <parameter key="rho" value="0.99"/>
  <parameter key="learning_rate" value="0.005"/>
  <parameter key="learning_rate_annealing" value="1.0E-6"/>
  <parameter key="learning_rate_decay" value="1.0"/>
  <parameter key="momentum_start" value="0.0"/>
  <parameter key="momentum_ramp" value="100000.0"/>
  <parameter key="momentum_stable" value="0.0"/>
  <parameter key="nesterov_accelerated_gradient" value="true"/>
  <parameter key="standardize" value="true"/>
  <parameter key="L1" value="1.0E-5"/>
  <parameter key="L2" value="0.0"/>
  <parameter key="max_w2" value="10.0"/>
  <parameter key="loss_function" value="CrossEntropy"/>
  <parameter key="distribution_function" value="bernoulli"/>
  <parameter key="early_stopping" value="false"/>
  <parameter key="stopping_rounds" value="1"/>
  <parameter key="stopping_metric" value="AUTO"/>

  <parameter key="stopping_tolerance" value="0.001"/>
  <parameter key="missing_values_handling" value="MeanImputation"/>
  <parameter key="max_runtime_seconds" value="0"/>
  <list key="expert_parameters"/>
  <list key="expert_parameters_"/>
</operator>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="136">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="514" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve TEST" from_port="output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Retrieve TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Deep Learning" to_port="training set"/>
<connect from_op="Deep Learning" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 2"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink result 1" spacing="0"/>
<portSpacing port="sink result 2" spacing="0"/>
<portSpacing port="sink result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure C.1: Experiments with DL H2O (Feature Set 1, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../../data/SMOTE TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kinoid|follower_count|in_reply_to_userid|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="112" y="187">
        <parameter key="repository_entry" value="../../../../data/TEST"/>
      </operator>
    </process>
  </operator>
</process>

```

```
<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="silent"/>
    <parameter key="random seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../../data/TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute filter type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time|user_id|following_count|account_creation_date"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value matrix row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning (2)" width="90" x="313" y="34">
```

```

<parameter key="activation" value="Tanh"/>
<enumeration key="hidden_layer_sizes">
  <parameter key="hidden_layer_sizes" value="50"/>
  <parameter key="hidden_layer_sizes" value="50"/>
</enumeration>
<enumeration key="hidden_dropout_ratios"/>
<parameter key="reproducible (uses 1 thread)" value="false"/>
<parameter key="use_local_random_seed" value="false"/>
<parameter key="local_random_seed" value="1992"/>
<parameter key="epochs" value="20.0"/>
<parameter key="compute_variable importances" value="false"/>
<parameter key="train_samples_per_iteration" value="16"/>
<parameter key="adaptive_rate" value="true"/>
<parameter key="epsilon" value="1.0E-8"/>
<parameter key="rho" value="0.99"/>
<parameter key="learning_rate" value="0.005"/>
<parameter key="learning_rate annealing" value="1.0E-6"/>
<parameter key="learning_rate_decay" value="1.0"/>
<parameter key="momentum_start" value="0.0"/>
<parameter key="momentum_ramp" value="1000000.0"/>
<parameter key="momentum_stable" value="0.0"/>
<parameter key="nesterov_accelerated_gradient" value="true"/>
<parameter key="standardize" value="true"/>
<parameter key="L1" value="1.0E-5"/>
<parameter key="L2" value="0.0"/>
<parameter key="max_w2" value="10.0"/>
<parameter key="loss_function" value="CrossEntropy"/>
<parameter key="distribution_function" value="bernoulli"/>
<parameter key="early_stopping" value="false"/>
<parameter key="stopping_rounds" value="1"/>
<parameter key="stopping_metric" value="AUTO"/>
<parameter key="stopping_tolerance" value="0.001"/>
<parameter key="missing values handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>

```

```

<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="112" y="187">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="false" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning" width="90" x="514" y="340">
  <parameter key="activation" value="Rectifier"/>
  <enumeration key="hidden_layer_sizes">
    <parameter key="hidden_layer_sizes" value="50"/>
    <parameter key="hidden_layer_sizes" value="50"/>
  </enumeration>
  <enumeration key="hidden_dropout_ratios"/>
  <parameter key="reproducible (uses 1 thread)" value="false"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="epochs" value="10.0"/>
  <parameter key="compute variable importances" value="false"/>
  <parameter key="train samples per iteration" value="-2"/>
  <parameter key="adaptive_rate" value="true"/>
  <parameter key="epsilon" value="1.0E-8"/>
  <parameter key="rho" value="0.99"/>
  <parameter key="learning_rate" value="0.005"/>
  <parameter key="learning_rate annealing" value="1.0E-6"/>
  <parameter key="learning_rate_decay" value="1.0"/>
  <parameter key="momentum_start" value="0.0"/>
  <parameter key="momentum_ramp" value="1000000.0"/>
  <parameter key="momentum_stable" value="0.0"/>
  <parameter key="nesterov accelerated gradient" value="true"/>
  <parameter key="standardize" value="true"/>
  <parameter key="L1" value="1.0E-5"/>
  <parameter key="L2" value="0.0"/>
  <parameter key="max_w2" value="10.0"/>
  <parameter key="loss_function" value="Automatic"/>
  <parameter key="distribution_function" value="AUTO"/>
  <parameter key="early_stopping" value="false"/>
  <parameter key="stopping_rounds" value="1"/>
  <parameter key="stopping_metric" value="AUTO"/>

```

```

<parameter key="stopping_tolerance" value="0.001"/>
<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="136">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="514" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Deep Learning (2)" to_port="training set"/>
<connect from_op="Deep Learning (2)" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 2"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<portSpacing port="source_input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure C.3: Experiments with DL H2O (Feature Set 2, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../data/SMOTE TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time|user_id|following_count|account_creation_date"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning (2)" width="90" x="313" y="34">

```

```

<parameter key="activation" value="Tanh"/>
<enumeration key="hidden_layer_sizes">
  <parameter key="hidden_layer_sizes" value="50"/>
  <parameter key="hidden_layer_sizes" value="50"/>
</enumeration>
<enumeration key="hidden_dropout_ratios">
<parameter key="reproducible (uses 1 thread)" value="false"/>
<parameter key="use_local_random_seed" value="false"/>
<parameter key="local_random_seed" value="1992"/>
<parameter key="epochs" value="20.0"/>
<parameter key="compute_variable_importances" value="false"/>
<parameter key="train_samples_per_iteration" value="16"/>
<parameter key="adaptive_rate" value="true"/>
<parameter key="epsilon" value="1.0E-8"/>
<parameter key="rho" value="0.99"/>
<parameter key="learning_rate" value="0.005"/>
<parameter key="learning_rate_annealing" value="1.0E-6"/>
<parameter key="learning_rate_decay" value="1.0"/>
<parameter key="momentum_start" value="0.0"/>
<parameter key="momentum_ramp" value="1000000.0"/>
<parameter key="momentum_stable" value="0.0"/>
<parameter key="nesterov_accelerated_gradient" value="true"/>
<parameter key="standardize" value="true"/>
<parameter key="L1" value="1.0E-5"/>
<parameter key="L2" value="0.0"/>
<parameter key="max_w2" value="10.0"/>
<parameter key="loss_function" value="CrossEntropy"/>
<parameter key="distribution_function" value="bernoulli"/>
<parameter key="early_stopping" value="false"/>
<parameter key="stopping_rounds" value="1"/>
<parameter key="stopping_metric" value="AUTO"/>
<parameter key="stopping_tolerance" value="0.001"/>
<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters"/>
</operator>

```

```

<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="112" y="187">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="false" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning" width="90" x="514" y="340">
  <parameter key="activation" value="Rectifier"/>
  <enumeration key="hidden_layer_sizes">
    <parameter key="hidden_layer_sizes" value="50"/>
    <parameter key="hidden_layer_sizes" value="50"/>
  </enumeration>
  <enumeration key="hidden_dropout_ratios">
  <parameter key="reproducible (uses 1 thread)" value="false"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="epochs" value="10.0"/>
  <parameter key="compute_variable_importances" value="false"/>
  <parameter key="train_samples_per_iteration" value="-2"/>
  <parameter key="adaptive_rate" value="true"/>
  <parameter key="epsilon" value="1.0E-8"/>
  <parameter key="rho" value="0.99"/>
  <parameter key="learning_rate" value="0.005"/>
  <parameter key="learning_rate_annealing" value="1.0E-6"/>
  <parameter key="learning_rate_decay" value="1.0"/>
  <parameter key="momentum_start" value="0.0"/>
  <parameter key="momentum_ramp" value="1000000.0"/>
  <parameter key="momentum_stable" value="0.0"/>
  <parameter key="nesterov_accelerated_gradient" value="true"/>
  <parameter key="standardize" value="true"/>
  <parameter key="L1" value="1.0E-5"/>
  <parameter key="L2" value="0.0"/>
  <parameter key="max_w2" value="10.0"/>
  <parameter key="loss_function" value="Automatic"/>
  <parameter key="distribution_function" value="AUTO"/>
  <parameter key="early_stopping" value="false"/>
  <parameter key="stopping_rounds" value="1"/>
  <parameter key="stopping_metric" value="AUTO"/>

```

```

<parameter key="stopping_tolerance" value="0.001"/>
<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="136">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="514" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Deep Learning (2)" to_port="training set"/>
<connect from_op="Deep Learning (2)" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 2"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<portSpacing port="source_input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure C.4: Experiments with DL H2O (Feature Set 2, after SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../../data/TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kincaid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute_value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute_block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning (2)" width="90" x="313" y="34">
        <parameter key="activation" value="Tanh"/>

```

```

<enumeration key="hidden_layer_sizes">
  <parameter key="hidden_layer_sizes" value="50"/>
  <parameter key="hidden_layer_sizes" value="50"/>
</enumeration>
<enumeration key="hidden_dropout_ratios"/>
<parameter key="reproducible_(uses_1_thread)" value="false"/>
<parameter key="use_local_random_seed" value="false"/>
<parameter key="local_random_seed" value="1992"/>
<parameter key="epochs" value="20.0"/>
<parameter key="compute_variable_importances" value="false"/>
<parameter key="train_samples_per_iteration" value="16"/>
<parameter key="adaptive_rate" value="true"/>
<parameter key="epsilon" value="1.0E-8"/>
<parameter key="rho" value="0.99"/>
<parameter key="learning_rate" value="0.005"/>
<parameter key="learning_rate_annealing" value="1.0E-6"/>
<parameter key="learning_rate_decay" value="1.0"/>
<parameter key="momentum_start" value="0.0"/>
<parameter key="momentum_ramp" value="1000000.0"/>
<parameter key="momentum_stable" value="0.0"/>
<parameter key="nesterov_accelerated_gradient" value="true"/>
<parameter key="standardize" value="true"/>
<parameter key="L1" value="1.0E-5"/>
<parameter key="L2" value="0.0"/>
<parameter key="max_w2" value="10.0"/>
<parameter key="loss_function" value="CrossEntropy"/>
<parameter key="distribution_function" value="bernoulli"/>
<parameter key="early_stopping" value="false"/>
<parameter key="stopping_rounds" value="1"/>
<parameter key="stopping_metric" value="AUTO"/>
<parameter key="stopping_tolerance" value="0.001"/>
<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>

<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="112" y="187">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="false" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning" width="90" x="514" y="391">
  <parameter key="activation" value="Rectifier"/>
  <enumeration key="hidden_layer_sizes">
    <parameter key="hidden_layer_sizes" value="50"/>
    <parameter key="hidden_layer_sizes" value="50"/>
  </enumeration>
  <enumeration key="hidden_dropout_ratios"/>
  <parameter key="reproducible_(uses_1_thread)" value="false"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="epochs" value="10.0"/>
  <parameter key="compute_variable_importances" value="false"/>
  <parameter key="train_samples_per_iteration" value="-2"/>
  <parameter key="adaptive_rate" value="true"/>
  <parameter key="epsilon" value="1.0E-8"/>
  <parameter key="rho" value="0.99"/>
  <parameter key="learning_rate" value="0.005"/>
  <parameter key="learning_rate_annealing" value="1.0E-6"/>
  <parameter key="learning_rate_decay" value="1.0"/>
  <parameter key="momentum_start" value="0.0"/>
  <parameter key="momentum_ramp" value="1000000.0"/>
  <parameter key="momentum_stable" value="0.0"/>
  <parameter key="nesterov_accelerated_gradient" value="true"/>
  <parameter key="standardize" value="true"/>
  <parameter key="L1" value="1.0E-5"/>
  <parameter key="L2" value="0.0"/>
  <parameter key="max_w2" value="10.0"/>
  <parameter key="loss_function" value="Automatic"/>
  <parameter key="distribution_function" value="AUTO"/>
  <parameter key="early_stopping" value="false"/>
  <parameter key="stopping_rounds" value="1"/>
  <parameter key="stopping_metric" value="AUTO"/>
  <parameter key="stopping_tolerance" value="0.001"/>

```

```

<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="136">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="514" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Deep Learning (2)" to_port="training set"/>
<connect from_op="Deep Learning (2)" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 2"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<portSpacing port="source_input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure C.5: Experiments with DL H2O (Feature Set 3, prior to SMOTE).

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.8.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="9.8.000" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve SMOTE TRAINING" width="90" x="45" y="34">
        <parameter key="repository_entry" value="../../../../data/SMOTE TRAINING"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="9.8.000" expanded="true" height="82" name="Select Attributes" width="90" x="179" y="34">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value="Avg_Syllables|Avg_Words_in_Sentences|class|Flesh-Kinoid|follower_count|like_count|Num_Big_words|Num_Long_Sentences|Num_Sentences|Num_Short_Sentences|Num_Syllables|Num_Uris|Num_words|Rate_Adv_adj|retweet_count|tweet_text|tweet_time"/>
        <parameter key="use_except_expression" value="false"/>
        <parameter key="value_type" value="attribute value"/>
        <parameter key="use_value_type_exception" value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type" value="attribute block"/>
        <parameter key="use_block_type_exception" value="false"/>
        <parameter key="except_block_type" value="value matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes" value="false"/>
      </operator>
      <operator activated="true" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning (2)" width="90" x="313" y="34">
        <parameter key="activation" value="Tanh"/>

```

```

<enumeration key="hidden_layer_sizes">
  <parameter key="hidden_layer_sizes" value="50"/>
  <parameter key="hidden_layer_sizes" value="50"/>
</enumeration>
<enumeration key="hidden_dropout_ratios"/>
<parameter key="reproducible (uses 1 thread)" value="false"/>
<parameter key="use_local_random_seed" value="false"/>
<parameter key="local_random_seed" value="1992"/>
<parameter key="epochs" value="20.0"/>
<parameter key="compute_variable_importances" value="false"/>
<parameter key="train_samples_per_iteration" value="16"/>
<parameter key="adaptive_rate" value="true"/>
<parameter key="epsilon" value="1.0E-8"/>
<parameter key="rho" value="0.99"/>
<parameter key="learning_rate" value="0.005"/>
<parameter key="learning_rate_annealing" value="1.0E-6"/>
<parameter key="learning_rate_decay" value="1.0"/>
<parameter key="momentum_start" value="0.0"/>
<parameter key="momentum_ramp" value="1000000.0"/>
<parameter key="momentum_stable" value="0.0"/>
<parameter key="nesterov_accelerated_gradient" value="true"/>
<parameter key="standardize" value="true"/>
<parameter key="L1" value="1.0E-5"/>
<parameter key="L2" value="0.0"/>
<parameter key="max w2" value="10.0"/>
<parameter key="loss_function" value="CrossEntropy"/>
<parameter key="distribution_function" value="bernoulli"/>
<parameter key="early_stopping" value="false"/>
<parameter key="stopping_rounds" value="1"/>
<parameter key="stopping_metric" value="AUTO"/>
<parameter key="stopping_tolerance" value="0.001"/>
<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>

```

```

<operator activated="true" class="retrieve" compatibility="9.8.000" expanded="true" height="68" name="Retrieve TEST" width="90" x="112" y="187">
  <parameter key="repository_entry" value="../../../../data/TEST"/>
</operator>
<operator activated="false" class="h2o:deep_learning" compatibility="9.8.000" expanded="true" height="103" name="Deep Learning" width="90" x="514" y="391">
  <parameter key="activation" value="Rectifier"/>
  <enumeration key="hidden_layer_sizes">
    <parameter key="hidden_layer_sizes" value="50"/>
    <parameter key="hidden_layer_sizes" value="50"/>
  </enumeration>
  <enumeration key="hidden_dropout_ratios"/>
  <parameter key="reproducible (uses 1 thread)" value="false"/>
  <parameter key="use_local_random_seed" value="false"/>
  <parameter key="local_random_seed" value="1992"/>
  <parameter key="epochs" value="10.0"/>
  <parameter key="compute_variable_importances" value="false"/>
  <parameter key="train_samples_per_iteration" value="-2"/>
  <parameter key="adaptive_rate" value="true"/>
  <parameter key="epsilon" value="1.0E-8"/>
  <parameter key="rho" value="0.99"/>
  <parameter key="learning_rate" value="0.005"/>
  <parameter key="learning_rate_annealing" value="1.0E-6"/>
  <parameter key="learning_rate_decay" value="1.0"/>
  <parameter key="momentum_start" value="0.0"/>
  <parameter key="momentum_ramp" value="1000000.0"/>
  <parameter key="momentum_stable" value="0.0"/>
  <parameter key="nesterov_accelerated_gradient" value="true"/>
  <parameter key="standardize" value="true"/>
  <parameter key="L1" value="1.0E-5"/>
  <parameter key="L2" value="0.0"/>
  <parameter key="max w2" value="10.0"/>
  <parameter key="loss_function" value="Automatic"/>
  <parameter key="distribution_function" value="AUTO"/>
  <parameter key="early_stopping" value="false"/>
  <parameter key="stopping_rounds" value="1"/>
  <parameter key="stopping_metric" value="AUTO"/>

```

```

<parameter key="missing_values_handling" value="MeanImputation"/>
<parameter key="max_runtime_seconds" value="0"/>
<list key="expert_parameters"/>
<list key="expert_parameters_"/>
</operator>
<operator activated="true" class="apply_model" compatibility="9.8.000" expanded="true" height="82" name="Apply Model" width="90" x="447" y="136">
  <list key="application_parameters"/>
  <parameter key="create_view" value="false"/>
</operator>
<operator activated="true" class="performance" compatibility="9.8.000" expanded="true" height="82" name="Performance" width="90" x="514" y="34">
  <parameter key="use_example_weights" value="true"/>
</operator>
<connect from_op="Retrieve SMOTE TRAINING" from_port="output" to_op="Select Attributes" to_port="example set input"/>
<connect from_op="Select Attributes" from_port="example set output" to_op="Deep Learning (2)" to_port="training set"/>
<connect from_op="Deep Learning (2)" from_port="model" to_op="Apply Model" to_port="model"/>
<connect from_op="Retrieve TEST" from_port="output" to_op="Apply Model" to_port="unlabelled data"/>
<connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled data"/>
<connect from_op="Apply Model" from_port="model" to_port="result 2"/>
<connect from_op="Performance" from_port="performance" to_port="result 1"/>
<portSpacing port="source input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="0"/>
<portSpacing port="sink_result 2" spacing="0"/>
<portSpacing port="sink_result 3" spacing="0"/>
</process>
</operator>
</process>

```

Figure C.6: Experiments with DL H2O (Feature Set 3, after SMOTE).

Bibliography

- [1] A Abdullah, MJ Awan, MF Shehzad, and M Ashraf. Fake news classification bimodal using convolutional neural network and long short-term memory. *preprint*, 2020.
- [2] Arush Agarwal and Akhil Dixit. Fake news detection: An ensemble learning approach. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1178–1183. IEEE, 2020.
- [3] Taruna Agrawal, Rahul Gupta, and Shrikanth Narayanan. Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1045–1049. IEEE, 2017.
- [4] Shahid Alam and Abdulaziz Ravshanbekov. Sieving fake news from genuine: A synopsis. *arXiv preprint arXiv:1911.08516*, 2019.
- [5] Jalal S Alowibdi, Ugo A Buy, S Yu Philip, and Leon Stenneth. Detecting deception in online social networks. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 383–390. IEEE, 2014.
- [6] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.
- [7] Jairo L Alves, Leila Weitzel, Paulo Quaresma, Carlos E Cardoso, and Luan Cunha. Brazilian presidential elections in the era of misinformation: A machine learning approach to analyse fake news. In *Iberoamerican Congress on Pattern Recognition*, pages 72–84. Springer, 2019.
- [8] Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. *arXiv preprint arXiv:1802.09913*, 2018.

- [9] Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82, 2019.
- [10] Judee K Burgoon, J Pete Blair, Tiantian Qin, and Jay F Nunamaker. Detecting deception through linguistic analysis. In *International Conference on Intelligence and Security Informatics*, pages 91–101. Springer, 2003.
- [11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [12] NLTK 3.4.5 Documentation. Natural Language Toolkit. www.nltk.org. [Online; accessed 19-January-2021].
- [13] Rapidminer Documentation. Deep Learning (H2O). docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/deep_learning.html. [Online; accessed 27-January-2021].
- [14] Rapidminer Documentation. Generalized Linear Model (H2O). docs.rapidminer.com/latest/studio/operators/modeling/predictive/functions/generalized_linear_model.html. [Online; accessed 27-January-2021].
- [15] Rapidminer Documentation. Naive Bayes (RapidMiner Studio Core). docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html. [Online; accessed 18-January-2021].
- [16] Rapidminer Documentation. Random Forest (Concurrency). docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_random_forest.html. [Online; accessed 18-January-2021].
- [17] Xishuang Dong, Uboho Victor, Shanta Chowdhury, and Lijun Qian. Deep two-path semi-supervised learning for fake news detection. *arXiv preprint arXiv:1906.05659*, 2019.
- [18] Hyungjin Gill and Hernando Rojas. Chatting in a mobile chamber: effects of instant messenger use on tolerance toward political misinformation among south koreans. *Asian Journal of Communication*, 30(6):470–493, 2020.

- [19] Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, Xiao-Li Li, and Guangxia Li. Deceptive review spam detection via exploiting task relatedness and unlabeled data. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1817–1826, 2016.
- [20] Wenlin Han and Varshil Mehta. Fake news detection in social networks using machine learning and deep learning: Performance evaluation. In *2019 IEEE International Conference on Industrial Internet (ICII)*, pages 375–380. IEEE, 2019.
- [21] Ionian University HILab, Department of Informatics. Fake News Detection For Hong Kong Tweets Data Set. hilab.di.ionio.gr/wp-content/uploads/2020/02/HILab-Fake_News_Detection_For_Hong_Kong_Tweets.xlsx. [Online; accessed 27-January-2021].
- [22] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 77–86. Springer, 2006.
- [23] Caio Libanio Melo Jeronimo, Leandro Balby Marinho, Claudio EC Campelo, Adriano Veloso, and Allan Sales da Costa Melo. Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 15–24, 2019.
- [24] Brendan Juba and Hai S Le. Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4039–4048, 2019.
- [25] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [26] Keras Keras. Deep learning library for theano and tensorflow. 2015, 2019.
- [27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [28] Abhishek Koirala. Covid-19 fake news classification with deep learning. *preprint*, 2020.

- [29] Marius Köppel, Alexander Segner, Martin Wagener, Lukas Pensel, Andreas Karwath, and Stefan Kramer. Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 237–252. Springer, 2019.
- [30] Sachin Kumar, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreti, and Mohammad Akbar. Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767, 2020.
- [31] Twitter Developer Labs. Twitter Standard Search API. developer.twitter.com/en/docs/twitter-api/tweets/search/introduction. [Online; accessed 25-January-2021].
- [32] Cheng Li, Xiaoxiao Guo, and Qiaozhu Mei. Deep memory networks for attitude identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 671–680, 2017.
- [33] Tim Loughran and Bill McDonald. Loughran-McDonald Dictionary. sraf.nd.edu/textual-analysis/resources/. [Online; accessed 18-January-2021].
- [34] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.
- [35] Mohammad Mahyoob, Jeehaan Al-Garaady, and Musaad Alrahaili. Linguistic-based detection of fake news in social media. *Forthcoming, International Journal of English Linguistics*, 11(1), 2020.
- [36] Adline Rajasenah Merryton and Gethsiyal Augasta. A survey on recent advances in machine learning techniques for fake news detection. *Test Engineering and Management*, 83:11572–11582, 2020.
- [37] P Read Montague. Reinforcement learning: an introduction, by sutton, rs and barto, ag. *Trends in cognitive sciences*, 3(9):360, 1999.
- [38] Despoina Mouratidis, Katia Lida Kermanidis, and Vilelmini Sosoni. Innovative deep neural network fusion for pairwise translation evaluation. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 76–87. Springer, 2020.

- [39] Despoina Mouratidis, Maria Nefeli Nikiforos, and Katia Lida Kermanidis. Deep Learning for Fake News Detection in a Pairwise Textual Input Schema (under review). *Computation (MDPI)*, 2021.
- [40] Maria Nefeli Nikiforos, Spiridon Vergis, Andreana Styliadou, Nikolaos Augoustis, Katia Lida Kermanidis, and Manolis Maragoudakis. Fake News Detection Regarding the Hong Kong Events from Tweets. In I. Maglogiannis, L. Iliadis, and E. Pimenidis, editors, *Artificial Intelligence Applications and Innovations. AIAI 2020 IFIP WG 12.5 International Workshops. AIAI 2020. IFIP Advances in Information and Communication Technology*, volume 585, pages 177–186, Cham, 6 2020. Springer.
- [41] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [42] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.
- [43] KBR Perera. The misinformation era: Review on deep learning approach to fake news detection. *preprint*, 2020.
- [44] Verónica Pérez-Rosas and Rada Mihalcea. Experiments in open domain deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1120–1125, 2015.
- [45] Rapidminer. RapidMiner Studio. rapidminer.com/get-started/. [Online; accessed 25-January-2021].
- [46] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012.
- [47] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

- [48] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach. Second Edition*. Prentice Hall, 2003.
- [49] Twitter Safety. Information operations directed at Hong Kong. blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html. [Online; accessed 25-January-2021].
- [50] Giuseppe Sansonetti, Fabio Gasparetti, Giuseppe D’Aniello, and Alessandro Micarelli. Unreliable users detection in social media: Deep learning techniques for automatic detection. *IEEE Access*, 2020.
- [51] Somayeh Shojaee, Masrah Azrifah Azmi Murad, Azreen Bin Azman, Nur-fadhlin Mohd Sharef, and Samaneh Nadali. Detecting deceptive reviews using lexical and syntactic features. In *2013 13th International Conference on Intelligent Systems Design and Applications*, pages 53–58. IEEE, 2013.
- [52] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020.
- [53] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
- [54] Panida Songram, Atchara Choopool, Paitoon Thipsanthia, and Veera Boonjing. Detecting thai messages leading to deception on facebook. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 293–304. Springer, 2016.
- [55] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.
- [56] Twitter. Twitter. twitter.com. [Online; accessed 25-January-2021].
- [57] Martijn Van Otterlo and Marco Wiering. Reinforcement learning and markov decision processes. In *Reinforcement Learning*, pages 3–42. Springer, 2012.
- [58] Ubobo Victor. *Robust Semi-Supervised Learning for Fake News Detection*. PhD thesis, Prairie View A&M University, 2020.

- [59] Dinesh Kumar Vishwakarma and Chhavi Jain. Recent state-of-the-art of fake news detection: A review. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2020.
- [60] William Yang Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.
- [61] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 516–523, 2020.
- [62] Wikipedia contributors. Flesch–Kincaid readability tests. en.wikipedia.org/w/index.php?title=Flesch%E2%80%93Kincaid_readability_tests&oldid=931233970. [Online; accessed 19-January-2021].
- [63] Alexandros Zervopoulos, Aikaterini Georgia Alvanou, Konstantinos Bezas, Asterios Papamichail, Manolis Maragoudakis, and Katia Kermanidis. Hong kong protests: Using natural language processing for fake news detection on twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 408–419. Springer, 2020.