NTNU
Norwegian University of
Science and Technology

Automatic Detection of Fake News in Social Media using Contextual Information

Torstein Granskogen

Master of Science in Computer Science
Submission date: January 2018
Supervisor: Jon Atle Gulla, IDI

Norwegian University of Science and Technology
Department of Computer Science

NTNU – Trondheim
Norwegian University of
Science and Technology

Automatic detection of Fake News in Social Media using Contextual Information

Torstein Granskogen

TDT4900 - Computer Science, Master thesis
Submission date: January 2018
Supervisor: Jon Atle Gulla
Co-Supervisor: Özlem Özgöbek

Department of Computer and Information Science
Norwegian University of Science and Technology

Abstract

Misinformation has become an important part of society, especially with the increase in fake news. This thesis investigates how using contextual and network data may be used as a detection system for news articles or other information pieces. Either as a standalone system or part of a bigger, hybrid solution.

A series of experiments have been conducted to explore the validity of contextual information in structured data from Facebook. Two different algorithms have been used, Logistic Regression and Harmonic Boolean Label Crowdsourcing, achieving a diverse result set shedding light on strengths and weaknesses. Using two different datasets consisting of scientific and fake news sources ranging from 4200 to 15.500 posts in size, and up to 9.5 million users, results with over 90 % accuracy in classification in supervised training scenarios, consolidating previous results on both old and new datasets.

As a result, this thesis concludes with very promising results using contextual data only. This approach is still novel and needs more rigorous testing, but combining it with existing Natural Language Processing systems might yield better results then the current state of the art systems. A lot of work is still needed to be able to apply the methods on less structured data.

Sammendrag

Falsk informasjon har blitt en viktig del av dagens samfunn, spesielt med den økende graden av falske nyheter. Denne avhandlingen utforsker hvordan bruk av kontekstuell informasjon og nettverksdata kan brukes som et detekteringssystem for nyhetsartikler og andre informasjonskilder. Enten som et eget system eller i en integrert løsning.

En rekke eksperimenter har blitt gjennomført for å utforske validiteten til kontekstuell informasjon i strukturerte data fra Facebook. To forskjellige algoritmer har blitt brukt, Logistic Regression og Harmonic Boolean Label Crowdsourcing, til å oppnå et diversifisert resultatsett som belyser styrker og svakheter i metodikken. Ved bruk av to forskjellige datasett bestående av vitenskaplige og falske nyhetskilder i størrelsesordenen 4200 til 15,000 poster og opptil 9,5 millioner brukere, har vi oppnådd resultater med over 90 % nøyaktighet med forhåndsklassifiserte data. Dette konsoliderer tidligere resultater både på gamle og nye datasett.

Denne avhandlingen konkluderer med veldig lovende resultater ved kun bruk av nettverksdata. Denne tilnærmingen er fortsatt ny og trenger mer detaljert og sterkere testing, men ved å kombinerere denne metoden med eksisterende språkprosseseringssystemer vil den kunne gi like gode, om enn ikke bedre resultater enn de beste systemene. Det gjenstår fortsatt mye arbeid for å kunne anvende metodene herfra på ustrukturerte data, men er lovende med tanke på mengden tilgjengelig data.

Preface

This thesis is submitted to the Norwegian University of Science and Technology(NTNU) as a part of the course TDT4900 - Computer Science, Master thesis. The research is conducted at the Department of Computer and Information Services (IDI), and was supervised by Prof. Jon Atle Gulla and postdoc Özlem Özgöbek.

Acknowledgements

First and foremost, I would like to thank my supervisiors Jon Atle Gulla and Özlem Özgöbek for providing me with continous and invaluable feedback throughout the thesis. They have pushed me forward during the work and shown massive interest and helped me with everything from structure to brainstorming the different aspects of the thesis.

In addition to this, I would also like to thank Eugenio Tacchini for advice and feedback when gathering the data and choosing sources. Without it, the thesis would have halted at an early stage.

Contents

List of Figures

List of Tables

CHAPTER 1

Research Overview and Summary

This initial chapter will present the problem domain, as well as give the research context. Included in this chapter is the research goals and questions, which are the motivation for this project. It will also include the outline of the report.

1.1 Background and motivation

Fake news has been in the public eye since 2016[3], which can be seen in the popularity rise from Figure 1.1. They have been spread by prominent politicians, known media houses and through other sources such as social media and word of mouth. The impact has been felt by most, where the validity of news stories and claims have been challenged both politically and scientifically. The trustworthiness of news agencies have been heavily disputed, and the use of the "fake news" has turned into a shouting match about what points of view that are accepted by different people, and thus becoming filled with emotions instead of facts.

The amount of fake information is increasing and spreads to more and more topics, but overall, the more technical and complex a topic is, the harder it is to produce false claims and information for it. The fakes produced changes just as the normal news changes, and is often based on the same topics. For example, during the United States presidential election in 2016, massive amounts of political news was published and spread, and therefore the amount of politically loaded fakes were also increasing. Fake news is increasing, but the fight against it is also increasing, and the overall awareness about it and how to spot it as well. Tools are needed as they evolve, both to minimize, but also to combat it.

This change in definition and usage of fake news in the public eye is something that might lead to the misconception about what fake news

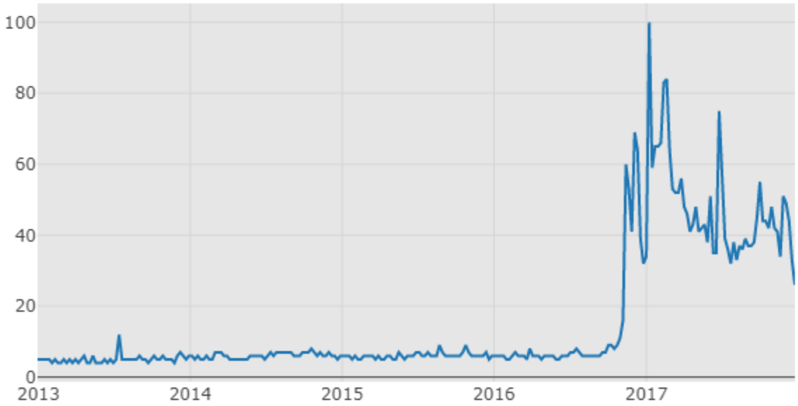

Figure 1.1: Fake news popularity on Google last 5 years

actually are. Because of this, the researchers working with fake news wants to change the wording of this kind of misinformation, as it entails so much more than just news, and it is not only news that can be fake. A more definite wording is needed to be able to discuss the different aspects better. This is discussed in detail in Chapter 2.

When looking at how people are accessing news[56][57], we can see that social media are one of the most prominent news sources in the U.S.[15]. Combining this with the fact that false information is spreading fastest on said social media sites, it is vital that the issue is taken seriously and either stopped or mitigated. It is spreading the fastest in topics and industries where there is a lot of emotion and points of view, where readers have a slight tendency to believe information that solidifies their existing views, even though they can be based on false information.

An example of how prominent misinformation in the media can be is during the U.S election in 2016, where more than 150 news stories where shared upwards of 40 million times on Facebook[3].

1.2 Problem Outline

When detecting and predicting the occurrence of false information, there are two main approaches. The most used approach is part of natural language processing(NLP), where the text itself is analyzed and based on the heuristics used. It is based on the understanding of the written language, and the method has been researched and improved upon since the Georgetown Experiment in the 1954[40]. NLP has adopted static approaches from N-grams to Support Vector Machines, as well as newer techniques involving machine learning and neural networks. The topic is discussed in greater detail in Chapter 2.

The other approach looks at the information around that is not directly a part of it, such as user data, sources, and network traffic. This approach incorporates the use of non-textual data and predicts based solely on those. This creates a universal approach where computers can be applied on languages where they do not yet have the adequate skill for normal linguistic methods. It is mainly a predictive technique, and cannot to the same extent as textual approaches decide whether or not information is truthful or not.

By using structured data from social media sites, specifically Facebook, we aim to create a graph that is used to determine the reliability of different users in the graph. The reliability is to what extent the user is either preferring false information or how active it is on sites or communities that are classified as spreading false information. This can then be extrapolated when incorporating more and more structured data by looking at how the nodes, the users, are connected through the edges of the social graph. The idea is to extend this approach to less structured data after the technique is matured, so that it can be used on news events, analyzing news sources and even being used by fact checkers as a tool aiding them in the textual evaluation of claims and facts.

The use of non-textual clues as the sole input for NLP is in an early phase, especially in the fake news domain. As a result, this thesis focuses on validation and further testing of the domain and idea. This involves testing the robustness of the method by testing it on new datasets with different density and sources.

The network data that have been used in the thesis originates from Facebook and is based on a collection of sources that are deemed either reliable or deceiving. It can also be seen as contextual data, as it is all bound to different information pieces, such as a post from a user or a community, whereas the reactions are the metadata that we are building our solution on. The thesis gives most attention to analyzing the existing data and how to best use it. Thus a series of experiments have been performed to figure out what data sets are the best to use, both in size and density. This is discussed further in 5.

1.3 Research Goals

Throughout the duration of this thesis testing the validity of contextual methods was paramount. It is a novel approach that need rigorous testing, and the research goals are affected by this.

- Can contextual data be used to successfully predict fake news in structured data on randomly chosen sources, or are they dependent on good, structured data?
- Does a web-of-trust augment prediction of fake news, and is it feasible to create one for bigger social networks?
- How volatile are contextual methods compared to each other, and how do their results change based on size, density and others measures?

1.4 Research Context

This thesis is conducted as a part of the course TDT4900 "Computer Science, Master thesis". The thesis as administrated and supervised by Prof. Jon Atle Gulla and co-supervised by Özlem Özgöbek. It is a part of the research project *Fake News Analytics at Scale (FNAS)*, which is a collaboration between industrial and educational actors. FNAS aims to develop technology that can be used to minimize the impact of false and malicious information in the news domain. The program aims to attack four areas:

- Classification of news against verifiable sources.
- Tracking of news origins.
- Network-based calculation of news sources credibility.
- Extraction and analysis of relevant facts of news.

This thesis falls under the third item, as presented in the Problem Outline, and is also a submodule of NTNUs SmartMedia Program[39]. The SmartMedia Program aims to develop architectures and technologies for large-scale real-time data processing. The intention of SmartMedia is to look into new technologies and how these might help journalists deal with new challenges.

1.5 Report Outline

The report is structured in a logical and standardized way. Chapter 2 gives a thorough theoretical background about what this thesis is about and what technologies and techniques are used. Chapter 3 is related work both in the same field, but also in neighboring fields that sheds light on important aspects of the thesis. Following this is chapter 4 which deals with the data that have been used for the experiments that are covered in chapter 5. Chapter 6 presents the results from the experiments given in chapter 5 and includes a discussion part about what the results mean and what can be extracted from them. Finally, chapter 7 is a combination of further work that can be done on the results from this thesis and a conclusion of what has been achieved.

CHAPTER 2

Theoretical background

This chapter will give an introduction to the theory that is used in this project, as well as examples on how some of these methods have been applied in some applications. It also tries to give a concise state of the art that is woven together with the theory, so that applications of the methods can be seen at once.

2.1 Fake news impact

The spreading of false information can be divided into many different groups based on what the intent or origin of the information. What most people think about when hearing the word "fake news" can to a certain degree be called propaganda. Propaganda is the conscious manipulation of people emotions and thoughts by using strong means and instruments to bring forth certain perceptions and actions. One example of this is to come forth with a claim that is outrageous. This claim will most likely be deemed untruthful later, but by bringing such claims, especially in politics, one can put other people in the limelight in a negative way, and thus change the way people think about them.

Furthermore, a somewhat prominent element of fake news lately are hidden paid posters, fake accounts and paid content on social media sites. For example in the net neutrality debate, evidence has come forth that millions of the comments used as evidence were fake comments[30] that abused real users names and the content was faked to help a certain view in the debate. There has been evidence pointing to news articles being made during the U.S election that was designed to promote one party's agenda. There were also articles for the other party, but the extent and magnitude were much lower. Additionally, there has sprung up areas where fake news and spreading of false information has become a business. There is a Macedonian[55] village where production of fake news has made people earn a lot of money. This kind of reaction is similar to the production of fakes in other industries. As long as there is a demand for the product, there will be someone to supply that.

Another aspect of fake news is simply false information. False information can be everything from wrongful facts, claims that are false to simple errors made during the creation of information or news stories. These fakes are harder to spot, since normally the entirety of the story is normally not created as a fake story, and such is a mix of truthful and wrong information. This might come from using outdated sources, biased sources to just making assumptions without checking the facts. This kind of "fakeness" is, for now, best handled by humans, as automating fact checking and validation of the correct claims and facts are hard for computers.

Another way to divide fake news is the intent of the information. This can first and foremost be divided into three parts, namely hoaxes, satire and malicious content. A hoax is a falsehood that fabricated to look like the truth. This can be events such as rumors, urban legends, pseudoscience. It can also be practical jokes, April Fools' Day jokes and so on. Hoaxes range from being in good faith, such as jokes, to malicious and dangerous stories such as pseudoscience and rumors. In addition to hoaxes, we have satire, in which something is ridiculed. Such as a public person being ridiculed in good faith where some of their more prominent sides are taken out of context and made even more visible. Satire can, as hoaxes, be both in good faith and for humor, but also be used in a malicious way to lower the standing of someone or something. Finally, we have content that is made with the intent of being destructive, namely malicious content. This content is made to destabilize situations, change public opinions and otherwise use false information to spread a message with the purpose to damage institutions, persons, political views or something similar.

One important thing to notice is that there is clear overlap between the different types of fakes. This happens because of the intent of the producer of the information. All of the different types mentioned above can be malicious if the wrong data is input in the information piece. The intent will be completely different. One example is papers or articles that have been published with wrong information that has been disproven at a later stage and continues to be used as a source by certain groups as proof of a point of view. This might then lead to splits in what is the correct science, where one can select what parts of sciences one want to believe in. This kind of destructive behavior is undermining the essence of empirical research and needs to be dealt with.

2.2 Detection of fake news

When working with fake news, propaganda or misinformation, it is always important to decide what one wants to detect. It is easier to have a specialized system that targets only certain parts of fake news. Then the system will have increased insight within its own domain, but will not be able to detect anything else. General systems that aim to detect several aspects are not as accurate because the system naturally has to adapt to changes and the rules cannot be as strict as in a specialized system.

An approach that works on contextual data has some advantages over language approaches. The contextual information is the same every time, regardless of language. Because of this, a contextual system can be applied to different datasets spanning languages with little to no changes. However, contextual approaches are not as definite as NLP approaches. Because they look at the probabilistic and statistical data, they will have a harder time detecting outliers, such as users that follow false information but do not spread or in any other form take part in the misinformation. They will to a contextual system look like a user that prefers false information, whereas the opposite is true. By analyzing the textual information, these details are detected. This shows how the different approaches, at all levels, are only as good as the techniques they use.

Other than how the detection or prediction is done, results are almost always better if the system is designed to detect parts of the spectrum. Targeting clickbait, satire, fake web pages and fake advertisements are examples of this kind of segmenting. Different approaches are better at certain problems. On the top level, we have linguistic and network methods, which work on different types of data. Linguistic methods analyze the language only, while network methods take into account the information surrounding the language, such as network traffic, user relationships, and links. These are explored further in following sections, together with general methods that can be used together with both linguistic and network methods to augment the results.

2.3 General approaches

The techniques and methods mentioned in this section can be used in both linguistic and contextual methods. They are methods that decide based on learning instead of rules, and therefore are able to improve based on the input over time, compared to static methods. One of the methods that have been used with success is machine learning. It has been used to train classifiers to improve the decision making of mostly linguistic methods, but can also be used for contextual and network-based methods. Deep learning has also been used together with several different techniques, mostly as neural networks.

To increase the efficiency of linguistic systems, machine learning and artificial intelligence approaches have been incorporated. This has made the systems even more resilient, and able to handle more and more generalized content because of the learning part that machine learning and AI has, compared to the more static early systems. In the two following sections, machine learning and artificial neural networks are presented in more detail.

2.3.1 Artificial neural networks

Neural networks is in short a programming paradigm where computers are enabled to learn from observational data, and thereby increase in efficiency and accuracy over time[58].

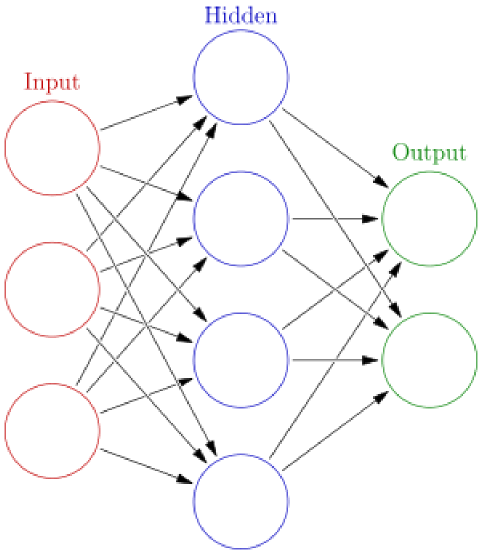

Figure 2.1: Artificial neural network, example network[63]

An Artificial neural network(ANN) is a system which imitates how the biological nervous systems process information, such as the brain. The brain itself is a series of interconnected neurons, which each works on solving the same problem, and through the learning potential is able to work out a better solution given more input data and time. ANNs are mostly configured against a single application, where it can specialize its learning potential within a single topic and the more specialized it becomes, the better it becomes to find data that does not fit the model it creates, and outliers and other abnormalities are then more easily detected. This can, for instance, be used in tax analysis, where systems can process the normal tax forms, but the moment something is not within the thresholds it can notify a human to have a look at it.

The main difference between neural networks and conventional computing is that ANNs do not follow a set path of instructions to find a solution, but instead organically finds a solution and therefore can be unpredictable if not given the correct training and input data.

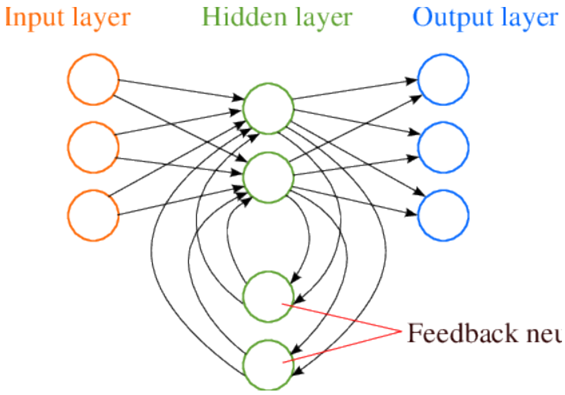

Figure 2.2: Recurrent neural network[23]

The neurons in ANNs are created where they fire an output based on certain inputs. They are also designed to operate in two modes, learning mode and use mode. Learning mode is used when training the neuron to fire to a given input. Since each of the neurons look at a minuscule part of the issue, one can combine many of them to look at for instance a picture, and if enough neurons fire on the given input, it can decide that there is a face in the picture, because of certain neurons have fired in a certain way and found their patterns.

Neurons can be much more complicated than the ones stated above. They can have weighted inputs, where certain inputs take precedence over other and will fire if the total input if over a threshold. The networks come in many different forms. Feed-forward networks, see Figure 2.1, always work in one direction, from input to output, and are mostly used in pattern recognition. Feedback, see Figure 2.2 networks can have signals travel back and forward in the network, and contain loops. Because of this, the state of the entire network is always changing, and will only give an output when the system is in a stable state. More complex ANNs use neurons that are called perceptrons, which are neurons with weighted inputs with some additional, fixed, pre-processing.

2.3.2 Machine learning

Certain tasks are extremely hard to program by hand and are better solved by machines. Face recognition, data mining, robot motion and other complex tasks often involve too many variables for a person to keep track of. Then it is better to adopt a computer to solve them instead, by utilizing the learning aspect of them. Machine learning can be used to solve complex tasks and relies on real-world data instead of intuition. This makes it slightly different from ANNs[33].

For machine learning(ML) training, there are three main approaches. Supervised learning is learning where the learning data have both inputs and outputs so that we always know the correct answer to the input, and can train and adapt the ML algorithm to get the same output as the correct one. The second approach is unsupervised learning, where some of the data only contain input. Because of this, the system makes some assumptions and thus unsupervised classifies the input without the known correct answer. Finally, reinforcement learning is learning where there is no direct access to the correct output, but the quality of the output can be measured following input. Reinforcement learning uses rewards to quantify the output and over time the model is changed based on how the total reward changes. This kind of heuristic approach where the model changes over time is similar to how ANNs work, and there is overlap between the methods, and also where they are applied.

2.4 Linguistic approach

The linguistic or textual approach to detecting false information involves using techniques that analyzes frequency, usage, and patterns in the text. Using this gives the ability to find similarities that comply to usage that is known in types of text, such as for fake news, which have a language that is similar to satire and will contain more emotional and an easier language than articles have on the same topic.

Below a selection of different prominent linguistic approaches has been detailed further. They are a selection of old, proven methods and some newer state of the art methods.

2.4.1 Support Vector Machines

A support vector machine(SVM) is a classifier that works by separating a hyperplane(n-dimensional space) containing input. It is based on statistical learning theory[59]. Given labeled training data, the algorithm outputs an optimal hyperplane which classifies new examples. The optimal hyperplane is calculated by finding the divider that minimizes the noise sensitivity and maximizes the generalization and margin of the model[41]. A unique feature of the SVM is that the hyperplane approach is based solely on the data points, and these points are called the support vectors. One of the major drawbacks with SVM is that it can only work with labeled data, and thus only work in a supervised training fashion.

SVM is not bound to linear separation, as they are able to transform input data into a high dimensional feature space, whereas a separating hyperplane can be found to work as an optimal classifier. One of the strengths of SVMs are that they can be used for very high dimensional problems, as long as their features can be mapped linearly in the feature space. The non-linear use of SVMs utilizes something called the *kernel trick*. The kernel trick works by replacing parts of the original algorithms with a kernel function instead of a dot function. Kernel methods can work in high-dimensional spaces because they compute the inner products between the data in the space instead of using the coordinates of the data. It is also worth mentioning that higher dimensional feature spaces increase the generalization error, but given enough samples, it still performs well.

2.4.2 Naive Bayes

Naive Bayes is a family of linear classifiers that works by using mutually independent features in a dataset for classification[46]. It is known for being easy to implement, being robust, fast and accurate. They are widely used for classification tasks, such as diagnosis of diseases and spam filtering in E-mail. If it is used on systems where the features are strongly dependent on each other, the performance normally takes a hit.

$$posteriorprobability = \frac{conditionalprobability * priorprobability}{evidence}$$

Naive Bayes is based on the probability rule of Bayes, which is shown above and can be interpreted as the probability of an object belonging to a class given the features it possesses. In addition to recognizing patterns, Naive Bayes can also be used in text classification by representing the text in a series of features. Naive Bayes classifiers are being used in many different fields, including diagnosis of diseases and decision making regarding treatment, the classification of RNA sequences in taxonomic studies and spam filtering in e-mail clients[46].

The naive part of Naive Bayes comes from the assumption that the variables are independent and identically distributed. This means that the variables used in the classification are all drawn from similar probability distributions. Independence means that probability of one outcome does not affect any other outcomes. Coin tossing is a good example of an independent and identically distributed collection. One outcome does not affect the other, and both variables have equal probability distribution.

2.4.3 Term frequency inverse document frequency

Term frequency-inverse document frequency(TF-IDF) is a weight value often used in information retrieval and gives a statistical measure to evaluate the importance of a word in a document collection or a corpus. Basically, the importance of a word increases proportionally with how many times it appears in a document, but is offset by the frequency of the word in the collection or corpus. Thus a word that appears all the time will have a low impact score, while other less used words will have a greater value associated with them[28].

$$TF(t) = \frac{Numberoftimestermtappearsindocument}{Totalnumberoftermsinthedocument}$$

$$IDF(t) = \frac{\log_{\epsilon}(Totalnumberofdocuments)}{Numberofdocumentswithtermtinit}$$

$$TF - IDF = TF * IDF$$

The term frequency is how frequent a term is in a document. A document here is a single piece of information, being a Facebook posts, a Twitter message or even a news article. The frequency will often increase in longer documents, and is normally divided by the document length if the collection consists of varying sized documents, as a way to normalize the values.

Finally, the inverse document frequency measures the importance of a term. While the term frequency does not discriminate between terms, the IDF part knows that words that occur often normally bring little quality to the document collection, and weighs these down while rare terms are scaled up. These widely used words can be stop words, words gotten from using stemming and other pre-processing tools.

2.4.4 N-grams

N-grams are a *n* long character slice of a longer string, but can also refer to n words. An n-gram with value one is called a unigram, two is bigram, three is tri-gram and so on. N-grams can be used to divide texts into different parts, and by using white spaces, it can also figure out what are words and not. N-grams are partially based on Zipf's Law, which can be stated as follows:

> The *n*th most common word in a human language text occurs with a frequency inversely proportional to *n*.

This means that some words will dominate other words, and that n-grams will have the same distribution, and that documents will have the same distributions as well, and we can compare them[6].

N-grams have been used in many different applications since it was first proposed, including code detection[1], evaluation of summaries[32] and automatic evaluation of machine translation[14]. The power of N-grams as an NLP tool is almost unprecedented, and it can be used for almost any task involving text and where the text is in a domain where there is some sort of frequency distribution in that domain.

2.4.5 Sentiment analysis

Sentiment analysis is a language processing tool that aims to identify the underlying viewpoints in a text. It tries to classify the sentiment polarity, which is a measure of the text being positive or negative compared to something[42]. It can be against a given corpus with weights added to words or n-grams of words or even check if certain texts are for or against something. The latter requires a lot of specialized work to be usable, but it is absolutely possible. Sentiment analysis can utilize resources with deep linguistic knowledge about sentiment indicators, thereby building on the existing knowledge about language.

Sentiment analysis tries to extract the opinions from a text and thereby be used further to evaluate something. This something can, for instance, be recommender systems or editorial sites trying to create summaries or come up with recommendations for the users. Sentiment analysis is a hard task for computers, as they in a way have to get under the skin of the writer of a text. This requires understanding in the way a user uses certain words. Since people use words differently, and languages are widely different in grammar and syntax, sentiment analysis requires massive amounts of work to be usable, such as a corpus containing weights that include most of the words that are used in that language[35].

Sentiments can mostly be extracted from opinions, as they contain subjective information. The opinions are normally built up consisting of two key components: a target and a sentiment on the target. The target can be any kind of entity, ranging from persons to events to a product. As such, one can say the the objective of sentiment analysis is to extract all opinions in a given text[35].

Sentiment analysis can be applied to most written knowledge, including but not limited to Facebook, Twitter, blogs, reviews, and discussions. It can also be used as a tool for organizations to gather public opinions beside surveys and polls. The raw data they have on different platforms can be used to generate a baseline through use of non-biased data. Domains where sentiment analysis have been utilized recently include healthcare, financial services, and political elections. Research on sentiment analysis includes prediction of sales performance, reviews to rank products, twitter sentiments compared to public opinions and many others[35].

2.5 Contextual approach

Contextual approaches incorporate most of the information that is not text. This includes data about users, such as comments, likes, re-tweets, shares and so on. It can also be information regarding the origin, both as who created it and where it was first published. This kind of information has a more predictive approach then linguistic, where you can be more deterministic. The contextual clues give a good indication of how the information is being used, and based on this assumptions can be made.

This approach relies on structured data to be able to make the assumptions, and because of that the usage area is for now limited to Social Media, because of the amount of information that is made public there. You have access to publishers, reactions, origin, shares and even age of the posts.

In addition to this, contextual systems are most often used to increase the quality of existing information and augment linguistic systems, by giving more information to work on for these systems, being reputation, trust metrics or other ways of giving indicators on whether the information is statistically leaning towards being fake or not.

Below a series of contextual methods are presented. They are a collection of state of the art methods and old, proven methods.

2.5.1 Logistic regression

Logistic Regression(LR) is a regression analysis that works when the dependent variable is binary. It is a predictive analysis and is used to explain the relationship between one dependent binary variable and other independent variables. Logistic regression can be used in situations where there is a yes or no question, such as whether or not a post on Facebook is considered fake or not. It can be seen as a special case of a linear model, and in the same family as linear regression. The main differences are that LR uses a Bernoulli distribution instead of a Gaussian distribution and that the outcome is a probability. LR will model the chance of an outcome based on the individual outcomes, and the result given will fall within the decision boundary. Using the decision boundaries, which decides what classification a result will receive, we get a result from the algorithm that is either *True* or *False*.

Logistic regression is based on the central mathematical concept the logit, the natural logarithm of an odds ratio. Logistic regression is well suited for describing relationships between categorical outcomes, such as classifying an information piece being false or not[44].

The simplest case of linear regression has only one predictor and one binary outcome variable. This can be formed like this:

$$logit(Y) = \log_{\epsilon}(\frac{\pi}{1-\pi}) = \alpha + \beta X$$

Where Y is the outcome of the binary variable, α is the Y intercept, β is the regression coefficient and X is the predictor.

Logistic regression has been widely used in social sciences for studying outcomes, such as promotions, divorce, medical diagnoses, unemployment and political voting.

Procedure 1 Regularized Harmonic Algorithm

```
Input: Bipartite graph $E \subseteq U \times I$, answers $a_{ui}$, $k_{max}$
Output: Estimation of correct solutions $s_i$ for all $i \in I$
for all $user\ u\ and\ item\ i$ do
    $\alpha_\mathrm{u} = 1 + \Delta > 0, and$
    $_\beta\mathrm{u} = {}_\alpha\mathrm{i} = {}_\beta\mathrm{i} = 1$
end for
for $k = 1..., k_{max}$ do
    for all user $u \in U$ do
        $p_u \leftarrow \alpha_u / (\alpha_u + \beta_u)$
    end for
    for all $item i \in I$ do
        $\alpha_\mathrm{i} \leftarrow 1 + \sum_{\mathrm{u} \in \vartheta \mathrm{i}} (\mathrm{a_{ui}}(2\mathrm{p_u} - 1))^+$
        $\beta_\mathrm{i} \leftarrow 1 + \sum_{\mathrm{u} \in \vartheta \mathrm{i}} (-\mathrm{a_{ui}}(2\mathrm{p_u} - 1))^+$
    end for
    for all item $i \in I$ do
        $p_i \leftarrow \alpha_i/(\alpha_i + \beta_i)$
    end for
    for all $user\ u \in U$ do
        $\alpha_\mathrm{u} \leftarrow 1 + \sum_{\mathrm{i} \in \vartheta \mathrm{u}} (\mathrm{a_{ui}}(2\mathrm{p_i} - 1))^+$
        $\beta_\mathrm{u} \leftarrow 1 + \sum_{\mathrm{i} \in \vartheta \mathrm{u}} (-\mathrm{a_{ui}}(2\mathrm{p_i} - 1))^+$
    end for
end for
Return estimate vector $\hat{s}_i = sign(\alpha_i - \beta_i) for\ all\ i \in I.$
```

2.5.2 Crowdsourcing algorithms

Crowdsourcing platforms often require analysis of their content to verify users or ideas. The verification can be seen as a binary labeling task[12], as the content is either verified or not. Crowdsourced tasks have human involvement, and the human factor often ends up in a fraction of the feedback being of poor quality, either due to malevolence or misunderstanding of the task. An example on this is the naming of the boat RRS Sir David Attenborough, where the public was allowed to vote and create names for a research vessel. The public majority wanted *Boaty McBoatface* as the name on the vessel, but the naming committee found it being a "*dilemma between credibility of the organization and burden of public opinion*". This is one of many examples where the public feedback is not on the same level as other parts and can be classified as poor quality feedback.

Two algorithms are proposed for the boolean labeling issue; The Regularized Harmonic Algorithm(RHA) and The Beta Shape Parameter Estimation Algorithm. The latter is not used in this thesis, but more information about it can be found in [12]. The RHA algorithm works by representing the knowledge about a user and the knowledge about an item both via beta distributions. Based on the shape parameters of the item, the user has an influence proportional to how close it is to the shape of the item.

This algorithm is used extensively in the experiments presented in Chapter 5, and the results are presented in Chapter 6 where it is called the Harmonic Boolean Label Crowdsourcing (HBLC) algorithm. The algorithm assumes that each member is associated with a certain quality score, y, as well as a probability, x, of telling the truth. Y is the actual probability of the member being perceived as true by a perfectly reliable user. The algorithm is better at distinguishing users of similar quality because it has information about the differing amount of certainty over them. Detailed information about RHA and its history can be found in [12] and [60].

2.5.3 Network analysis

Social network analysis is a powerful tool for working with graphs. Figuring out the connections and relationships between different vertices are important to be able to extract non-direct results from Facebook, Twitter or other network-based applications. The analysis can be used to predict behavioral patterns in a population that shares everything from communities, likes, friends and so on. The social behavior among people indicates that shared interests, in general, will lead to similar behavior, such as belief in fake news. The network perspective emphasizes the structure of the relations, instead of the attributes of the individual actors. The need for a statistically significant population is still paramount, and something that this thesis is looking into[51].

Network analysis has been successfully used to improve customer satisfaction[19], to help understand how learners collaborate[13], user attribute and behavior analysis, interaction analyses, link prediction and recommender system development to mention a few of the areas. In addition to this, it is applied on a human level concerning intelligence, counter-intelligence and law enforcement activities. Knowing how the population or certain types within a population reacts to a certain input is vital for reacting in a correct way. Equally, network analysis can be used on any collection of data where the items have a relationship, such as a text corpus, thus exploring word usage, word relationships and so on.

2.5.4 Trust Networks

Trust networks or webs of trust are a graph that differs from the original networks and offers a better view of the actual relationships between nodes. A trust network is a simple weighted directional graph where each connection between nodes holds a trust value. Trust networks are assumption based, and will only provide a best-effort approximation on how trustworthy a user thinks another user is. Furthermore, one can use propagation to expand the trust network from non-direct connections, for example through transitivity or structural similarities.

The network is openly available, which means that all the nodes have access to and can read the information that the other nodes contain. The trust values can differ based on the usage of the graph, and besides the normal 0 to 1 it can also have distrust values where the total range is -1 to 1. The network can also be segmented and subsets are taken out of it, giving trust values within smaller sets.

One of the main advantages with trust networks is that they are able to make qualified assumptions based on the information they have, normally their individual trust networks and they do not take into account hearsay and other non-fact based assumptions. A few things that are important to mention is that a lack of trust does not automatically mean distrust. A lack of a trust measure can also mean that there is little to no interaction between the different actors, and therefore they have no trust values. These values can as stated earlier be propagated, but sometimes that is not a good option. Another important factor is that actual trust is not the same as likes on Facebook or upvotes on Reddit, but instead is based on the underlying interactions and relations. Trust itself shows a confidence in that the particular node contributes to what another node believes in.

Trust Metrics

Trust metrics are the quantification of trust between two nodes. A good trust metric focuses on supporting the capture of values in a reliable and standardized way. A metric like this is the spreading activation models from Cognitive Science, which is designed for computing subjective neighborhoods of most trustworthy peers in the network[61].

Several pseudo-trust metrics exists on major websites today, including eBays Feedback rating, Slashdots karma, Reddit's upvote/downvote system and also Wikipedias reputation scores. Most of these are not true trust metrics, as they do not create the score independent of user input, but they are examples on how content can be ranked based on the quality and relevance of it. They are a push in the right direction, but trust networks require a considerable amount of structured data, and systems today were not created with this in mind. They require substantial amounts of work to integrate such metrics.

2.5.5 Content-driven reputation system

Reputation systems work by giving users the ability to rate each other based on their performance. The reputation systems range from sitewide reputations like on eBay and Amazon, to review sites where each review has its own reputation independent on the user. The reputation systems help other users to find real-world feedback on sellers and wares.

One of the main issues with reputation systems is that they are based on subjective feedback. It is extremely hard for people to be fully objective, and as such the reputation that is given to a product, a seller, a person or any other entity will be colored based on the opinions of others instead of objective measures. For a user-driven community like Wikipedia, where all the content is created by people for people, it is essential that the content is both correct and holds a certain standard. If the content quality lessens over time, users will notice and the site might lose reputation and users.

To combat this, it is possible to have an automated content-driven reputation system that objectively measures the trust and reputation of authors. Authors are evaluated based on their contribution to the page. Instead of users giving other users thumbs up or down, by leaving edits done by authors, they are implicitly giving them the vote of confidence and builds the original authors trust, edit by edit. Based on the authors keeping the edits of other authors, the reputation will change according to the reputation of the authors reviewing it. This way, authors or users cannot damage the reputation of others by simply inserting negative comments or ratings. Instead, they have to remove the edit made, then replace or change it to something else, with the risk of other authors restoring the original edit, and then all the work is wasted[2].

The notion of reputation has good predictive value, where changes performed by low-reputation authors have a higher probability of poor quality than high-reputation authors. Even so, as with many NLP tasks, the reputation system is hindered by the lack of understanding the markup language that was used in [2], and thus the text analysis is of poorer quality then it should be. It should also be mentioned that reputation systems can be classified into two categories: chronological, where the reputation is calculated from the chronological sequence of ratings, and fixpoint, where the reputation is calculated over the entirety of the feedback, without any consideration of temporal information. The reputation system used by eBay is a chronological one, while PageRank is a fixpoint reputation system.

2.5.6 Knowledge Graphs

Knowledge graphs can be seen as an extension of the original search using keywords only. It enables the search engine and other knowledge-based entities to build up a knowledge base that contains certain important aspects of entities, such as the Taj Mahal. The knowledge graph consists of many nodes connected where each node contains some information, for example, geographical location, and then other nodes are connected to that either directly by having that relation, or on a different level, like country. The knowledge graph that Google uses for their search, augments the search that users do with information that is relevant[54]. That kind of graph is similar to the use of linked data, ontologies and RDF, which all are ways of representing relationships within data in one way or another. They all try to make the relationships are connections between entities computer-readable.

Enabling computers to understand the connections between entities enables systems to drastically increase the usability of information stashes, as they now can use downtime between use to extract more information from the existing data. In addition to finding connection within stored data, knowledge graphs can use other sources, such as the CIA World Factbook, Wikidata, and DBPedia to augment their existing knowledge. An example illustration of a knowledge graph can be seen in figure 2.3, where it is clear that one person has connections to other persons, buildings, locations and other entities. By traversing this graph, and using it together with other information sources, it is possible to obtain a more detailed and complete understanding of entities.

2.6 Information extraction

Information extraction is the automation extraction of structured information, such as entities, relationships between entities and attributes describing entities from unstructured sources[52]. This is somewhat similar

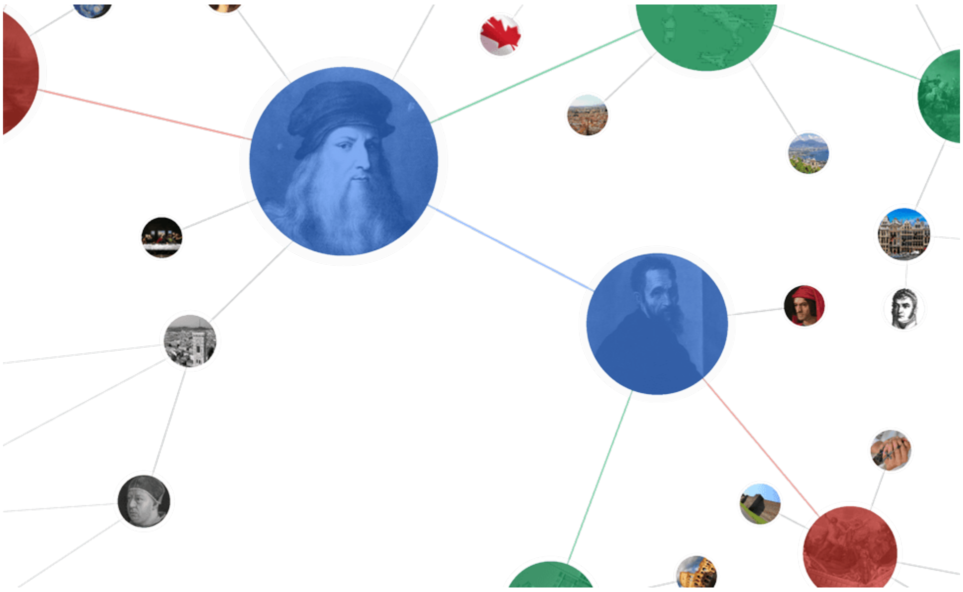

Figure 2.3: Conceptual illustration of a Knowledge Graph

to what knowledge graphs are for search engines. Information extraction includes structure extraction using machine learning, information retrieval, databases, web and document analysis.

Information extraction can be used in many different applications, such as news tracking. News tracking involves automatically tracking specific entities from news sources, such as an event, a person or even journalists. Data cleaning is a different application, where data warehouses need to keep their data in certain formats and them being compatible with each other. With information coming from a series of sources, the data needs to be understood, maybe transformed, and finally inserted into the warehouse for further storage and processing, even analysis.

2.6.1 Source extraction

Source extraction involves the extraction of the origin of data, or from what kind of source the information is gathered from. Getting the source origin is not a topic that has gotten much interest from the research community, as most of the work has been related to the NLP part of extraction, where the information is much more hands-on and is always available and is definite. In a contextual manner, the source can be used as an origin where information can be gotten further, for example, if one has a news agency listed as a source, it is possible to create a knowledge graph from them based on the traffic they receive. Most of the source extraction work has been done on the body and the headlines of texts and relies more on entities, keywords, and events.

Source extraction by type of source can be divided into structured and unstructured sources. Structured sources already have a logical order, and can be handily extracted. Unstructured sources, however, can vary in degree[52]. They are often augmented with existing structured sources, like databases, already labeled unstructured text or use of knowledge libraries. The most popular form of unstructured data are small text snippets or records, as their simplicity and minimal length minimizes the complexity. Other unstructured sources are paragraphs and documents, which are often needed to get the bigger picture in a text and to understand the context of the text items.

For more ambiguous sources, the accuracy of the extractor is impacted heavily by the homogeneity of the style and format in the document. Machine generated pages are often heavily templatized, like HTML and XML documents. On the other end of the spectrum are open-ended sources. Open-ended sources are among the hardest sources to work with because they are not within any given domain, such as medicine or astronomy, and therefore it is hard to make a system that can make sense of the contents easily. Sources like the Internet is open-ended, and when extracting information from it, there is a high possibility that redundant information is extracted but extracted differently[52].

2.6.2 Entity extraction

Entity extraction is when the extraction system is able to extract information that is related to an entity[52]. An entity is any type of thing that can have relationships, traits and other information stored with it. This can be persons, locations, companies or products, in addition to pretty much anything else. Entity extraction can also be addressed by segmenting a text record into structured entities, as a bridge between unstructured and structured data.

Named-entity extraction is the task where names are recognized in the text, such as persons, organizations, and locations. These entities are explicitly named in the text, and are easier to grab than more abstract entities that might be hidden in the text, and has seen a lot of research. This includes rule-based algorithms and machine learning systems, that utilizes a plethora of different feature and evaluation methods. One of the challenges in named-entity extraction is recognizing one entity is the same as another. For instance, *the automotive company created by Henry*

Ford in 1903 can be referred to as both *Ford* and as *Ford Motor Company*[38]. By including ontologies or other knowledge systems this can be minimized, but still remains a challenge, especially in open-ended domains, where redundant and similarly named entities can be issues as well.

2.6.3 Keyword extraction

Keyword extraction is an important part of information extraction, especially in the search domain, where understanding the essentials of a text is paramount. It is the automatic identification of a set of the terms to best describe the subject of a document or collection. Extracting the correct keywords will give a much better accuracy and relevance for a search engine, and also decide what documents are related to each other[36].

TF-IDF is one of the algorithms that have been widely used in keyword extraction, as it will highlight the frequency of the words normalized to the corpus in a ranked fashion. It does not require much input and is relatively domain-independent. TF-IDF is an example of a largely unsupervised method for keyword extraction but supervised and semi-supervised methods are also available.

Another well-known keyword extraction method is TextRank[37]. TextRank is a graph-based ranking model for text processing. The TextRank keyword extraction is based on unsupervised training and is based on *co-occurence* relations. First, the text is tokenized and annotated with part of speech tags. Only single words are considered as candidates for addition to the TextRank graph. Then all the lexical units that pass the syntactic filter are added to the graph, and those that co-occur within a window of N words, and the score is decided by ranking them over 20-30 iterations. In the post-processing part, all the candidates for keywords are collapsed into multi-word keywords.

CHAPTER 3

Related Work

Fake news has gotten a massive amount of exposure since 2016, and it is still important in the public eye. Because of this, the research into how it is impacting the society and how it can be battled has increased. This chapter presents some of the most prominent research done concerning fake news, misinformation, and propaganda.

3.1 Current state of fake news detection research

Most of the detection based research in fake news detection has been done on the textual level using well known Natural Language Processing(NLP) tools. This includes use of Support Vector Machines[8][11] and Naive Bayes[11][8] as classifiers for widely different issues. Clickbait[8] has been detected with good results using a mix of methods, including, but not limited to frequency analysis, neural network analysis, and image detection, and shows the strength of combining different methods based on the input.

Detection of something as specific as fake news, clickbait or hoaxes can be seen as a subset of NLP research. Because of this, it is also ridden by the same issues. One of the main issues with NLP is that many of the methods are only really good in closed domains, where the context and average user is already known. Both linguistic methods and network analysis have shown high accuracy within these closed domains[11], but for something as vague as fake news, which spans many topics and domains, a more robust system is needed. Systems that work on the contextual clues[60] might be better suited in the open domains, as the user patterns and relationships do not change hugely based on the domain. [60] proposes a detection based system based solely on the users and not related to the contents at all. In highly structured settings like social networks or linked data, it is possible to achieve really good results, even with simple methods like Logistic Regression.

In addition to detection and classification of fake news into the categories "Fake" or "Non-Fake", systems created to understand the context of the textual cues have shown good results. Fake News Challenge[21] (FNC) was a competition aimed at Stance Detection. Stance detection is determining the relative perspective a news source takes towards a specific claim. A detailed summary about FNC can be found in section 3.3.1. The competitors used widely different methods, but the best results were achieved using mixtures of modern and well-tested methods. The winning team used a mixture of deep learning and decision trees[64], where the tree model included well-known methods like TF-IDF, Word2Vec and sentiment features. The deep learning model consisted of a convolutional, feed-forward ANN. Other high ranked participants also used neural networks[4][47], in addition to other methods like Latent Dirichlet Allocation[4], Cosine Similarity[47] and TF-IDF[10].

Other researchers have taken a more generalized approach regarding the issue at hand and aims at understanding it at a more theoretical level. This includes type classification of fakes[50], where fakes are divided into serious fabrications, large-scale hoaxes, and humorous fakes. The categories are overlapping, but the level of maliciousness is what decides. [11] looks at the previous work done and shows that a hybrid approach using both linguistic and network methods is the most promising. It also sheds light on the purpose of fake news detection systems. They should be created to augment the human judgment, not replace it. Another important finding is that classifiers that are training sufficiently are able to detect instances of deception based on clustering. Hoaxes and fakes will not fit into a well-trained model. Equally important it sheds light on what the language in fake information looks like, where it is emotional, judgmental and exaggerates more than other articles on the same topic. [60] suggests that the user base of a structural community can be used to indicate whether or not information is credible, or at least give an indication. People tend to have a slight preference for truth, which can be spotted in big enough populations.

The concept of trust is another part of fake news that has been researched[61][49]. The actual interactions between users can be extracted into a sparse network[61]. This is based on the user activity and influence among each other, and not based on subjective information at all. It is similar to the content-driven reputation system that is proposed for Wikipedia[2]. Trust in this matter is the subconscious interactions and one of the most objective measures we currently have in social interactions regarding social networks. It can be seen as binary, discrete and continuous values based on the type of interaction extracted.

Additionally, a study has shown that as much as 80% of troll or hoax information on Facebook are from users usually interacting with the content already[5]. This is an indicator that populations will tend towards what they already believe in and reinforces the claims by [60].

3.2 Impact research

Some call the era which we are in now information-wise as the post-truth era[48], where facts and evidence have been replaced by belief and emotion. They state that the current social and political climate makes the current efforts of classification and objectively state a source as true or fake is not viable. The structure of voting schemes like Facebook only facilitates positive feedback, whereas the negative feedback is missing, and exposure to false information can only rise but not be buried. The issue with fake news is that the definition is not set in stone, and it is becoming more and more normal that fake news are just news that personal beliefs go against, whereas real "fake news" are deliberately created with false information to look like real news.

Consequently, others state that the current news environment incentivizes speed and spectacle[9], instead of investigative journalism that many of the established news agencies are built upon. The line between user-generated content and traditional news is increasingly blurred, and as a consequence of this, quality and truth is the first casualty in the war of income. Combining this with the issue of hidden paid posters[7], where paid users are generating content for hidden purposes. If done correctly, these users can change the opinion of topics and the political landscape over time.

3.3 Existing fact-checking and fake news detection systems

To challenge the increase of false information, a few systems worth mentioned have sprung up to deal with the issue. They show usage of methods and ideas presented in this thesis, and range from pure fact-checking sites that look at claims and facts in different news stories, to sites that aim to detect fake news in a broader setting. Factmata[17] is one of these. Factmata is a fact-checking community that is leveraged by artificial intelligence, and their goals are to reduce online misinformation and to help journalists, media enthusiasts, as well as advertisers and businesses. ClaimBuster[26] aims to be an automated, live fact-checking platform. It monitors live streams, websites and social media to catch factual claims. It was used to fact-check claims during the 2016 U.S Presidential Debates, to verify the claims used by the participants as close to real-time as possible. Another well-known fact-checking site is PolitiFact[45], a former winner of the Pulitzer Prize. They are a fact-checking website that rates the accuracy of claims by elected officials in American politics.

An interesting fact-checking organization that has surfaced in Norway is Faktisk[18]. Faktisk differs from many other similar systems as it is an ideal and independent organization that is a collaboration between most of the major news agencies in Norway. They work to check claims and facts that have been put forth in the public. These checks are done by professionals. An interesting fact about them is that when false claims have been made that they have checked, their version which includes a "rightness"-meter, tend to spread more quickly in social media than the original. This is an indication that readers are after the true stories, and that when what is actually true is known, it will let itself be known.

In addition to all these different systems that are already in production and being used to combat misinformation, it should be mentioned that several of the biggest information dealers on the web, such as Facebook, Google, Wikipedia, and Twitter are all actively fighting fake news on their platforms[31]. Google trying to monitor their searches as to not encourage fake hits being given to the end user, and Facebook removing posts that are deemed "hoaxes shared by spammers" for personal and monetary reasons. The sites are incorporating detection tools and letting users comment on content being fake or not. This is another indication that for the big information dealing companies to keep their position and be taken seriously, they have to do their part to keep the information clean.

3.3.1 Fake News Challenge

Fake News Challenge is a linguistic challenge that was put forth for participants to try out different approaches in a competitive setting. The goal was "to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem". It had a task to predict the stance of a news piece from a combination of a body and headline from a news article, either belonging to each other or not. Thus they were set to estimate the relative perspective of two pieces relative to a topic, claim or issue. They could go for agree or disagree, whether or not the body and headline conforms or not, or they could go for discussed or unrelated, if the body and headline discusses but not takes a position or does not discuss the topic at all.

Generally, the top results use artificial intelligence in one way or another to get their results, as presented in Section 3.1. Most of them are using multi-layered perceptrons together with well-known and tested features in linguistics.

3.4 Datasets

In addition to the dataset presented in Chapter 4, there are other datasets and sources that are worth mentioning. Other sources then the ones chosen from [24] can be used, for satire research or other purposes, even to see how fake news detection systems act on semi-fake news stories.

Liar[62] is a dataset consisting of almost 13.000 manually labeled short statements in various contexts from PolitiFact. This includes analysis reports and source documents for each case and is a good dataset for training linguistic systems.

Emergent[20] is another dataset containing concerning journalism. It contains 300 rumored claims and almost 2.600 associated news articles. The claims and news articles have been collected and labelled by journalists as **True**, **False** or **Unverified**. It also contained stance information on whether the article is **For**, **Against** or **Observing**. Emergent provides a dataset that is usable for linguistic systems, similar to Liar. The dataset used by Fake News Challenge[21] was based on Emergent.

Furthermore, datasets that contain contextual information such as user information, network data, images and source origins are very sparse, often because of the ambiguous terms of service that the sources present, and because of privacy concerns. Nevertheless, good sources for contextual data are social networks, such as Twitter and Facebook, which both have good platforms for gathering data.

CHAPTER 4

Data

The paper has used data from two different sets of data. These datasets are gathered using the same methods, and the only difference between them are the sources they are gathered from and the time period which the data are gathered from.

4.1 Some like it hoax

The first dataset is a recreation of a dataset used in [60]. They used a dataset gathered from public posts and posts' likes from selected Facebook pages. These were gathered in the time span between 2016-07-01 to 2016-12-27 and was gathered on 2017-01-27. Because of uncertainties on the terms of service for collection of data from Facebook using the Facebook Graph API[16], the data needed to be gathered again for this thesis. This was done on 2017-10-18. As mentioned in chapter 2, Facebook and other companies are taking steps to minimize the footprint of misinformation [24], and thus some of the original data gathered in [60] was not available at the data gathering. This has lead to the original and the newly extracted data set to have differences. The sources for the dataset are divided into two classes, namely scientific and non-scientific sources. The non-scientific pages are known to publish or embrace fake information, whereas the scientific ones are known to only publish truthful information. This leads to a two-way differentiation, where we have two major nodes that contain the extremes that help us in differentiating news stories. This clear-cut differentiation makes prediction and classification results much easier to read, as a blueprint is already made and comparisons can easily be made. The sources used can be found in Table 4.1.

The first time the data was gathered it consisted of 15.500 posts from 32 different pages (14 non-scientific and 18 scientific), with more than

2.300.000 likes by 900.236 users. Among the posts, 8.923 are hoaxes and 6.577 non-hoaxes. The second time the data set consisted of 4286 posts from 30 different pages (12 non-scientific and 18 scientific), with 418.476 likes by 158.789 users. Among the posts, 1389 are hoaxes and 2897 non-hoaxes. The main reason for the size difference is that data gathered from Facebook can be quite volatile. Some pages are taken down centrally by Facebook itself because of them not following the TOS, while others can be taken down by the posters themselves. Some pages stop existing altogether, are marked as private so that you need privileges to access them. All of these combined can and will lead to differences in large data gathering. Because of this, a direct comparison between the two datasets are not possible, but patterns and similarities can still be extracted.

Table 4.1: Sources for data set

Scientific	**Non-scientific**
Scientificast	Scienza di Confine
Cicap.org	CSSC - Cieli Senza Scie Chimiche
Oggiscienza.it	STOP ALLE SCIE CHIMICHE
Queryonline	vaccinibasta
Gravitazeroeu	Tanker Enemy
COELUM Astronomia	Scie Chimiche
MedBunker	MES Dittatore Europeo
In Difesa della Sperimentazione Animale	Lo sai
Italia Unita per la Scienza	AmbienteBio
Scienza Live	Eco(R)esistenza
La scienza come non l'avete mai vista	Curarsialnaturale
Liberascienza	La Resistenza
Scienze Naturali	Radical Bio
Perché vaccino	Fuori da Matrix
Le Scienze	Graviola Italia
Vera scienza	Signoraggio.it
Scienza in rete	Informare Per Resistere
Galileo, giornale di scienza e problemi globali	Sul Nuovo Ordine Mondiale
Scie Chimiche: Informazione Corretta	Avvistamenti e Contatti
Complottismo? No grazie	Umani in Divenire

4.2 New dataset

The second dataset used was gathered using the same methods as the first, but on different sources. These sources differed on multiple areas, most notably on community size and language. Whereas the original was limited to only Italian pages and communities, the new dataset is using several big, mainstream sources, such as The Washington Post and BBC News. The data was gathered from the same time period as the original so that the content skewing would be minimized, or at least be affected by the same actions, either random or deliberate. This dataset is divided in the same way as the other, with scientific and non-scientific sources only. The sources were chosen to check if locale, location or topics have a significant impact on the results or not. The size is comparable to the latest version of the original, but the contents are widely different.

With 5943 posts from 16 pages (7 non-scientific and 9 scientific), over 9.5 million likes and 5.6 million unique users, the dataset is much bigger in contents, even though the number of posts are close to a third of the original. Some of the sources chosen did not have any data and was not used in the gathering part. The posts were divided into 2558 hoaxes and 3385 non-hoaxes, so the ratio between types was close to the same, 43 % versus the original 31 %. The comparatively denser dataset might lead to different results from the original. The sources themselves can be found in Table 4.2.

The sources are based on two different sources. The majority of sources are gathered from [27] which contains a list of political sources that are divided into Real, Satire or Fake. The sources used in this thesis are the real and fake ones that contained data when extracting using Facebooks Graph API[16]. In addition to these sources, more known fake sources were gathered from the findings in [24]. It is a collaboration between Facebook and PolitiFact. They have classified a series of sites on Facebook as either parody, news imposter, fake news or duped sites. The sources used in this thesis was chosen randomly from the fake news category, resulting in the sources seen in Table 4.2.

4.3 Data usage

The data gathered contained the posts from the different sources of scientific and non-scientific news, together with the likes from the posts, including likes in comments. The likes were concatenated into post ID instead of keeping them linked to the individual comments. The posts were sorted into what community they belonged to, such that a hierarchy was created. The top level is the source, like *The Wall Street Journal*, followed by all of its posts that all contain the ID of those that have liked the content of each post. The source had its own unique ID, created by Facebook. Each post has an ID that in part contains the source ID fol-

Table 4.2: New dataset sources

Scientific	**Non-scientific**
The Wall Street Journal	Before it’s News
The Economist	InfoWars
BBC News	Real News. Right Now.
NPR	American Flavor
ABC News	World Politics Now
CBS	We Conservative
USA Today	Washington Feed
The Guardian	American People Network
NBC	Uspoln
The Washington Post	US INFO News
	Clash Daily

lowed by a post unique number. The only information stored and used is the unique ID number, no other information was needed, and thus not used.

CHAPTER 5
Experiments

This chapter consists of an overview of the experiments done, and provides details on what methods and choices were taken to get the results presented in chapter 6.

5.1 Comparison between datasets with the same sources

As mentioned in chapter 4, reproducing the original dataset from [60] was not possible to do fully, and a dataset of smaller size was created instead. These datasets originate from the same sources, and can, therefore, be used for comparisons. The size differences might lead to different results, because of the sparser dataset, and the user details might also skew the results somehow.

This experiment was performed by using the logistic regression and the harmonic boolean crowdsourcing algorithms on the newly created dataset and comparing them with the original results from [60].

Some changes had to be made to the algorithms for them to run fully during the tests. Some of the content from some sources were no longer available, and could therefore not be used in the experiment. In addition to this, because of the reduced size, when using logistic regression, the same granularity was not possible to obtain, as the sample size had shrunk quite significantly.

5.2 Comparison between different datasets

Since some of the results from [60] were astonishingly good, there was a need for another unrelated test using the same algorithms on a different dataset. This dataset was created to differ both in size and density, but with the same type of data. The sources were chosen without analyzing the contents, such that the results weren't skewed because of a clever choice of sources.

This dataset underwent the same experiments as the dataset that was recreated, such that the results could be compared and analyzed.

5.3 Different sizes on datasets

To test the robustness of the algorithms, the size of the datasets were changed to look for breakpoints in a viable solution.

5.3.1 Differing post amount

The new dataset was split a total of five times, leading to a series of new datasets.

- 1/2 size with randomly chosen content.
- 1/4 size with randomly chosen content.
- 1/8 size with randomly chosen content.
- 1/16 size with randomly chosen content.
- 1/32 size with randomly chosen content.

The split was made using reservoir sampling on the full dataset. Reservoir sampling is an algorithm family that randomly chooses k samples from a list of n items. It is mostly used when the data does not fit the memory constraints of a system[22].

The different sized datasets were then fed through the same algorithms as previously and then underwent the same analysis as the full sized one. In this experiment, the posts were randomly chosen, and they preserved all their information about users and likes, such that the user density and relationships were untouched.

CHAPTER 6

Results

This chapter presents the results gathered from the experiments detailed in chapter 5. They are presented in the same order as the experiments and contains discussion in each section. Some generalized results are presented in additional sections.

6.1 Comparison between same dataset

This section is for the experiment where the results from [60] are compared to the ones found using the regenerated dataset.

Table 6.1: Results from [60]

	One-page-out		Half-pages-out	
	Average accuracy	σ	Average accuracy	σ
Logistic Regression	0.794	0.303	0.716	0.143
Harmonic BLC	0.992	0.023	0.993	0.002

Table 6.2: Results regenerated dataset

	One-page-out		Half-pages-out	
	Average accuracy	σ	Average accuracy	σ
Logistic Regression	0.732	0.363	0.745	0.093
Harmonic BLC	0.978	0.075	0.955	0.062

In *One-page-out* all the posts from one page at a time are placed in the testing set, while the remaining posts from other pages are placed in the training set. In *Half-pages-out* a set consisting of half the pages are placed in the testing dataset, while the remaining half are placed in the training set.

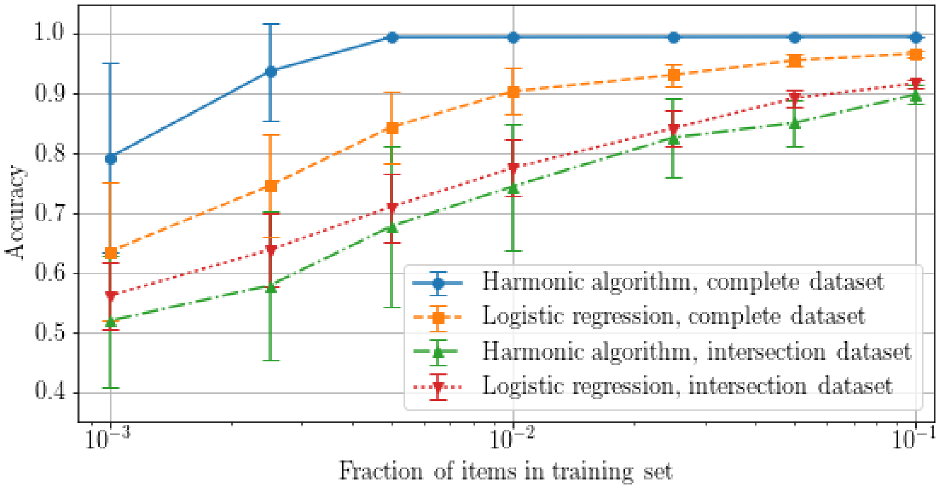

Figure 6.1: HBLC results, original dataset from [60]

We can see from the tables that the results are quite similar even though the regenerated dataset is about 30 % of the original size. This leads to showing that the robustness of both algorithms are able to handle smaller sizes and also perform with satisfactory results. We can still see that the results are overall worse for the smaller dataset, both with *One-page-out* and *Half-pages-out*.

Interestingly enough, even though the end results on the different dataset sizes are similar, the drop off were Harmonic Boolean label crowd sourcing(HBLC) starts losing accuracy because of items in the training set starts much earlier in the smaller, regenerated dataset, and that Logistic Regression(LR) is even able to outperform it when the size is small enough. This is an indication on that HBLC does need a certain amount of information to be a viable solution and is discussed further in the size comparison section.

HBLC performs overall better on the complete dataset, which contains all the users from every source. For *One-page* it scores close to perfect, with an accuracy of 97.8 % on the full set, and 95.5 % accuracy for *Half-page*. This is close to the results that [60] got in their experiments, and is a validation of the results gotten there.

For the intersection dataset, which contains all the users that have a connection to both the scientific and the non-scientific sources, HBLC performs much worse than Logistic Regression and is not much better than guessing. This is most likely because of the small number, 1360 out of 158.789. This is not a statistically big enough number for HBLC to be able

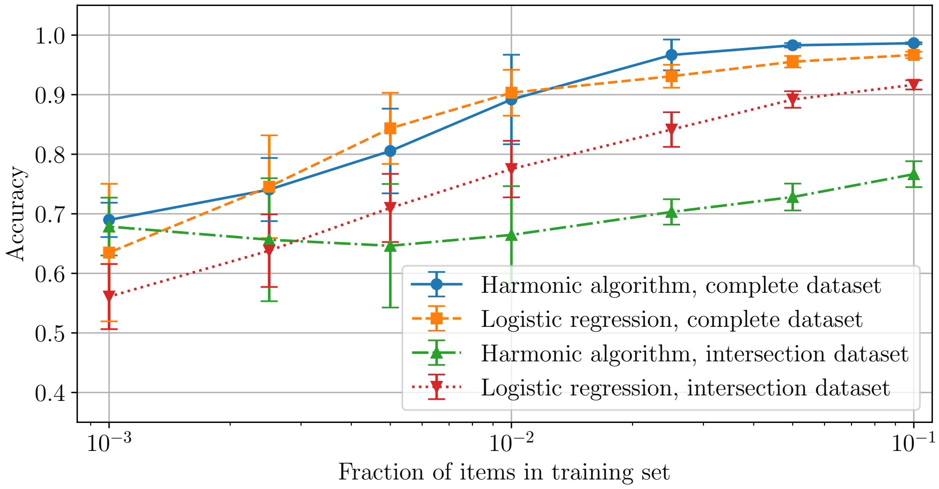

Figure 6.2: HBLC results, regenerated dataset

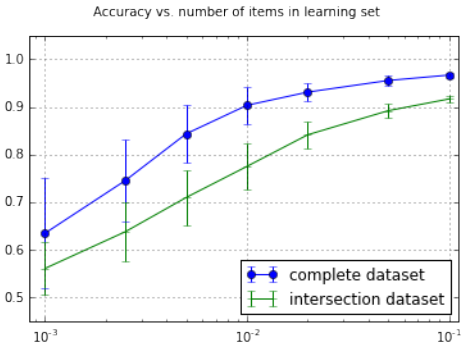

Figure 6.3: LR, original results

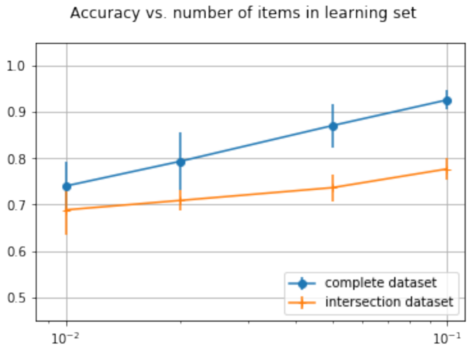

Figure 6.4: LR, regenerated results

to make sense out of the data, either with the training set being too small to find patterns or relationships between the users.

In addition to this, only 59.166 of the 158.789 users occurs more than once in the dataset. These users are hard to use for pattern recognition, as the information about them is just not there. These users, which can be called "dead" users, does not contribute enough to the end result.

6.2 Comparison between different datasets

This experiment looks at the differences between the regenerated dataset and the new dataset based on different sources.

The new dataset is much denser in the sense where it contains a lot more users than the regenerated one. The average like density is about 16 times higher, where the new dataset has an average of 1611 likes per post, the regenerated one has 97. The new dataset has an average of 1 like per user, whereas the regenerated has 2. The user density per post is 943 users per post for the new dataset, and 37 for the regenerated one. This is a big difference that is made deliberately to check how the different algorithms handle different inputs.

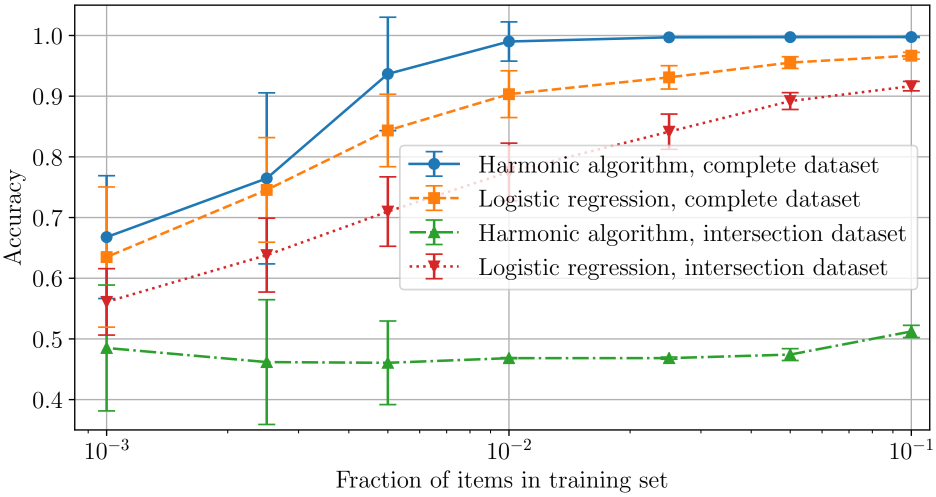

Figure 6.5: HBLC vs original LR, new dataset

Furthermore, the new dataset, while containing over 9 million unique users, only 14.616 of these are part of the intersection dataset, about 2 ‰of the total, whereas the regenerated had about 9 ‰. About 1.3 million users had liked more than one post, so the vast majority could be

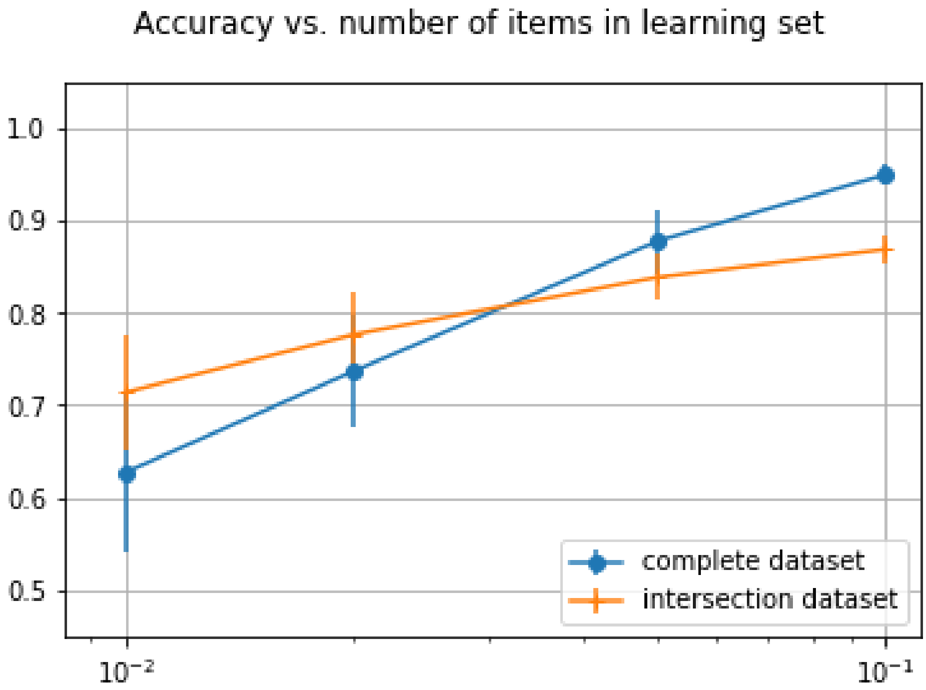

Figure 6.6: LR results, new dataset

classified as useless.

We can see when comparing Table 6.3 with Table 6.2 that the results from the regenerated dataset are overall better, but both are still performing well. The standard deviation increases significantly for HBLC, but not for LR. This indicates that HBLC relies more on the internal differences in the data than LR, and is also to be expected since HBLC is able to extract some relationship data that it can use for later calculations, whereas LR only uses the data available at all stages. Garbage in, garbage out is a well known saying where the quality of results rely on the input, and is most likely the case here as well, with the high fraction of "dead" users that the new dataset contains.

On the other hand, LR performs similarly for both the complete and intersection dataset, which indicates that it is a really robust algorithm that needs little input for it to find patterns. As seen in figure 6.6 there is an intersection point where the intersection dataset performs better than the complete dataset, but that is most likely to the fraction difference not fully giving reliable results because of the low starting number. This does not occur for the regenerated dataset, Figure 6.4, or the original, Figure 6.3, from [60].

Overall, the quality of the original dataset looks to be higher than the new dataset while the new dataset is probably more realistic. There will always be unworkable data, and the dead users can to a certain degree be compared to the cold start problem in recommender systems.

Table 6.3: Results new dataset

	One-page-out		Half-pages-out	
	Average accuracy	σ	Average accuracy	σ
Logistic Regression	0.772	0.288	0.683	0.121
Harmonic BLC	0.939	0.234	0.906	0.102

6.3 Different sizes on datasets

This section looks at the differences based on a changing size of the new dataset.

For this experiment, the new dataset containing 5943 posts was divided into smaller sizes. These sizes were halved each time, down to the final size which was 1/32. During the tests, because of the smaller sizes, it was not possible to run the LR tests on it without changes, and therefore the results are for HBLC only. Figure 6.5 is the results on the full dataset, while Figure 6.7, 6.8, 6.9, 6.10, and 6.11 are the divided datasets ranging from half to 1/32.

Table 6.4: Results new dataset, different sizes, complete dataset

	One-page-out		Half-pages-out	
	Average accuracy	σ	Average accuracy	σ
1/2	0.9379	0.2346	0.9299	0.0.0922
1/4	0.9354	0.2344	0.9352	0.0962
1/8	0.9323	0.2411	0.8730	0.1071
1/16	0.9112	0.2452	0.8698	0.0982
1/32	0.9162	0.2413	0.8559	0.1008

Looking at Table 6.4 and 6.5 we can see that gradually smaller sizes leads to larger standard deviation and a lower accuracy. Even though the smallest set is only 185 posts, HBLC manages a 90+ % accuracy using *One-page*. For *Half-page* the accuracy is 85 %, which is still good based on the size. Because of the bigger training set using half of the pages, the standard deviation is smaller.

For the intersection dataset, the results are quite abysmal. From the average accuracy using *One-page* on the full with only 35 %, to 8.7 % for

Table 6.5: Results new dataset, different sizes, intersection dataset

	One-page-out		Half-pages-out	
	Average accuracy	σ	Average accuracy	σ
1/2	0.2823	N/A	0.3829	0.1243
1/4	0.2276	N/A	0.3132	0.1132
1/8	0.1901	N/A	0.3487	0.1140
1/16	0.1246	N/A	0.3161	0.0974
1/32	0.0872	N/A	0.2951	0.1020

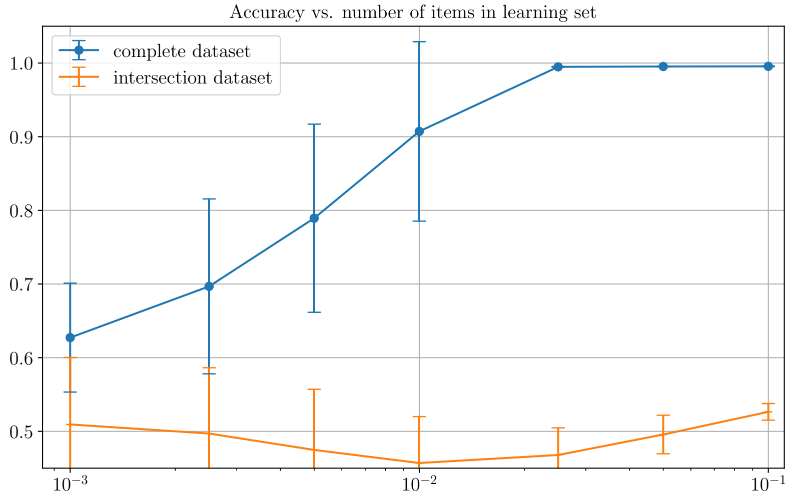

Figure 6.7: HBLC results, half size.

the smallest dataset. These data are based on a small amount of users, but interestingly *Half-page* shows more promise on small datasets, even though the accuracy is not good enough, with 42 % on the full, down to 29.5 % on the smallest. This indicates that for smaller datasets, using HBLC needs a certain amount of training data to be able to make sense of the input. It would be interesting to gather a much larger dataset consisting of only intersection users and perform the tests on them, too see how the algorithms perform with conflicting data.

Looking at the figures where differences in the fraction of items in the dataset are tested, for all the sizes, there is a significant drop off in the complete dataset accuracy based on the size. As the dataset gets smaller, this drop off occurs closer to the full size of the training set, and based on the graphs, a minimum of 1.000 posts looks to be the viable amount of HBLC.

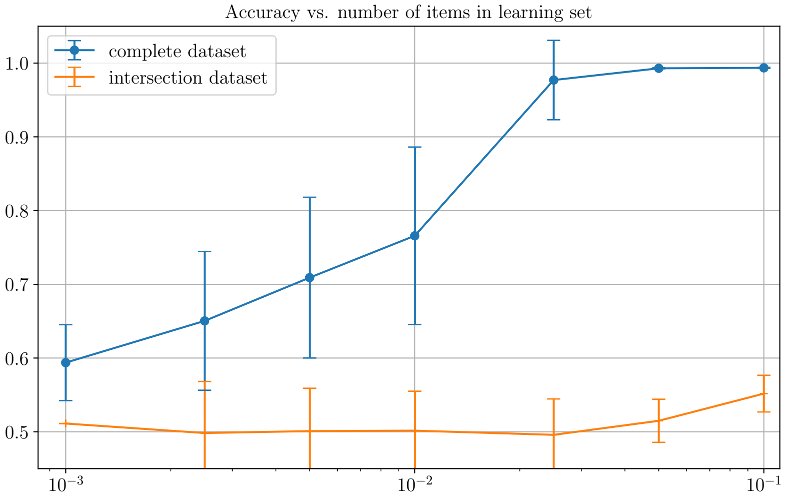

Figure 6.8: HBLC results, quarter size.

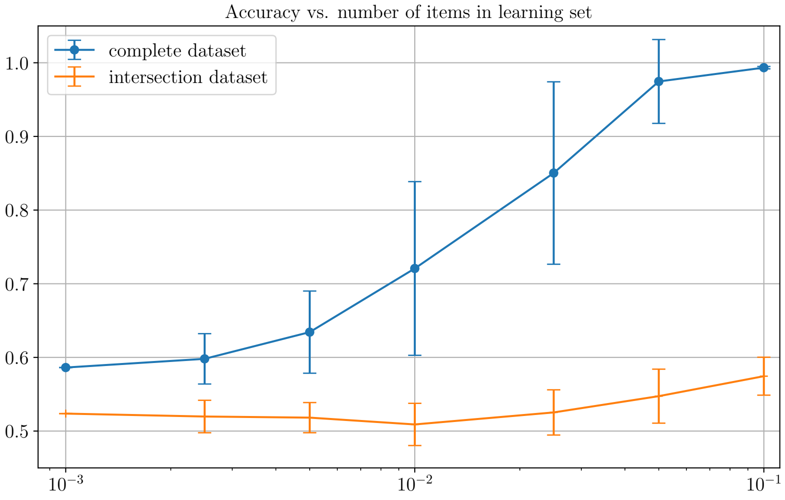

Figure 6.9: HBLC results, 1/8 size.

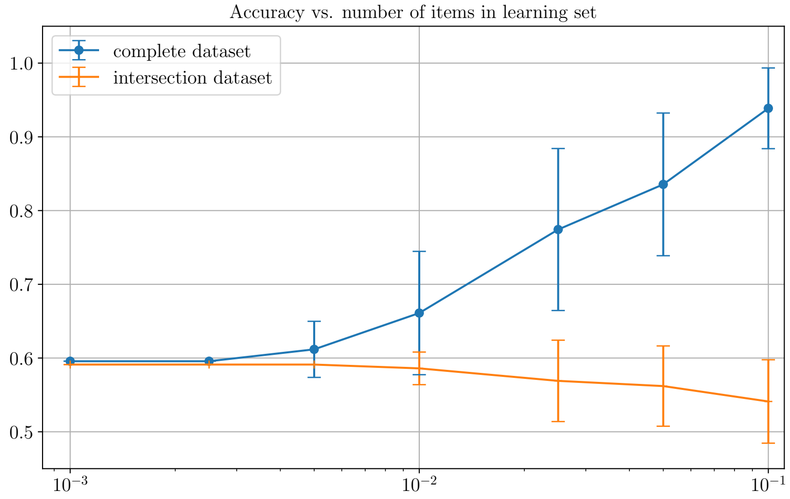

Figure 6.10: HBLC results, 1/16 size

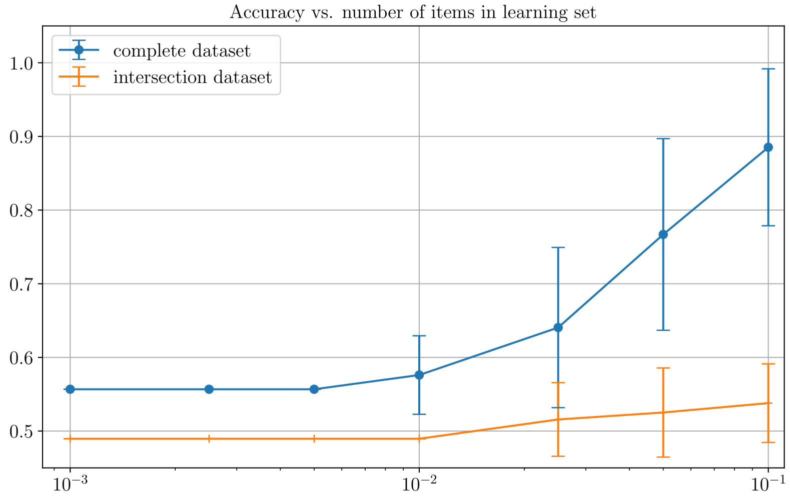

Figure 6.11: HBLC results, 1/32 size

It's worth noting that the density of the dataset was maintained during the split, with the average of likes per post ranging from 1611 down to 1485 for the smallest split. More details about the dataset sizes can be found in Table 6.6.

Table 6.6: New dataset, size details

Dataset	#Posts	#Likes	#Users	#I.Users	#IU 2+	#U 2+	Likes/Post
full	5943	9.576.262	5.646.218	14.616	5372	1.336.069	1611
1/2	2971	4.641.242	3.101.897	6896	2948	629.404	1562
1/4	1485	2.484.998	1.940.333	2578	1148	272.834	1673
1/8	742	1.235.440	1.032.672	1479	888	116.644	1665
1/16	371	556.742	488.188	228	126	44.954	1500
1/32	185	274.907	255.107	132	82	14.443	1485

6.4 Comparison between the algorithms

The two algorithms used in this thesis, Logistic Regression and Harmonic Boolean label crowdsourcing, are both good candidates for getting good results on network data. LR is the most stable algorithm and performs similarly on the different datasets. With an accuracy of over 70 % on all datasets using the full training sets, it is a good result, but not great.

HBLC is a more advanced algorithm, and this shows in the results. It performs with an accuracy over 90 % on all the different datasets and shows a lot of promise for further tests on network data. It is to a degree able to deduce patterns in the data and this enables it to perform at such a high level. Despite this, it relies both on more and better quality of data than LR.

Overall, HBLC is the better algorithm to use for this kind of problem, as it consistently outperforms the simpler LR. If a small or sparse dataset is used, LR can perform close to HBLC, but with complex and dense datasets that often are available in social networks, HBLC will make the cut.

6.5 Additional results

After the previous experiments were completed, it was possible to extract some more information from the data and is discussed in this section.

One of the major issues with the raw data from the Facebook Graph API is that so many of the users in the data are low quality, meaning that they contribute very little. They contain just one or a few likes in the posts gathered, and can therefore be seen as "dead" or low-quality users. Because of these users, which measures upwards of 80-95 & of the total, the complexity of the calculations skyrockets when working with large amounts of data.

Closely related to this is the intersection users, which are users with likes in both scientific and non-scientific posts, amount to between 2 ‰and 9 ‰. As mentioned, it would be interesting to test the algorithms against a bigger set (hundreds of thousands) to see how it handles conflicting data on its own.

This can be mitigated with several methods. One of them is performing preprocessing on the data after gathering, where the total user contribution is calculated, and only users of a certain quality are used. The quality heuristic needs to be explored further, but can be so simple as a like amount threshold.

Furthermore, the source selection is important to the results, as the hoax and non-hoax sites need to be within the same multitude or close to it for it not to be enough training data to make proper comparisons. Mixed pages that are not part of the extremes need to be explored, where the algorithms need to classify data themselves. This is a vital part of making a mature fake news detection system.

CHAPTER 7

Conclusion

This chapter will conclude the findings in during this project, and also shows some preliminary results to the research questions, together with the further work that is both planned and needed for a successful approach of this mainly theoretical paper.

7.1 Discussion

The thesis has undertaken a novel and challenging task of evaluating the usability of contextual data as the only source in the detection of false information in structured networks. The results in the previous chapter have shown that the results are promising, but that they are volatile and that more research is needed to be able to more precisely classify the contextual approach as viable or not. The experiments are done on simple sources that are extremes on the fake news scale, and the following section aims to present the strengths and weaknesses of the system.

7.1.1 Evaluation

The way this thesis has detected fake news is novel, and similar results have been hard to find. Most detection of fake news has been done using linguistic cues. Results can be compared with how good they are on similar datasets, but the methods are different, and therefore not fully comparable. The results are similar to the ones achieved in [60], which is to be assumed since they are based on the same algorithms, but on different datasets. They show a lot of promise, and the results are really good, achieving over 90 % accuracy on both datasets, and might very well be part of a tool that can be used in a bigger scale later. Even though the results are good, as mentioned by others[11][8] a hybrid solution will most likely yield the best and most stable results. One way is to use contextual clues as starting data that linguistic systems can use later on.

However, the results are only based on a small subset of sources and needs to be rigorously tested on both bigger and more diverse sets. The methods need to be tested against data that contains more than only hoaxes or scientific parts. In addition to this, the results gotten from the algorithms are ambiguous, in as it is hard to read whether or not the posts, the users or the sources are the most critical part of getting good results. It is not possible to read a single reason to why the results change the way they do when shrinking the dataset size.

7.1.2 Data quality

The data gathered were processed raw, which means that no further work was done it before any of the tests were conducted. This can be seen in the results, where, amongst others, a vast amount of the users had a really low activity count in the dataset. It is really hard to predict the usage patterns of such users, and therefore they do not contribute enough to justify being part of the calculations. In addition to this, the post density was vastly different between the sources. The bigger non-fake sources contained many posts with a huge amount of users, whereas the users in the fake sources were fewer, but the average user had better quality. These results are not the same as was shown during the U.S. presidential election where fake posts on average had more coverage. The choice of sources were deliberate to check the robustness and veracity of the algorithms, but a more standardized dataset is needed to be able to evaluate results against each other, especially when the data gathered are quite volatile in a way where two datasets from the same sources will not contain the exact same contents if they weren't taken within a time span of a month or sometimes even less. The results showed that the algorithms were robust enough to handle different input and still have almost the same results.

Furthermore, the distribution of user types was very skewed towards certain users. If dividing the users into three groups; intersection users, low-quality users and high-quality users, the majority was low quality, and a minuscule amount were intersection users. The sources should be chosen such that the algorithms could be tested against a more diversified dataset. The results from such datasets will give much clearer answers towards the validity of the algorithms against the vast majority of the sources that are not yet classified.

7.1.3 Technical challenges

The experiments in this thesis did at times take very long to conclude, upwards of 18 hours using a moderately powerful desktop PC. The complexity of the algorithms increases exponentially with more data. As a result, the new dataset and its almost 10 million users took a long time to compute. When combining this with the number of low quality users, it is paramount that preprocessing of a kind is needed for an efficient and viable solution. One way to achieve this could be to apply a best-effort solution, like simulated annealing. Simulated annealing is a probabilistic technique for finding the global maximum of a heuristic. If given a time constraint, it will give the best answer found within that time frame. For a real-time application, this is vital, and the solution can work at a later time in finding the actual global maximum.

7.2 Concluding remarks

After exploring the fake news domain regarding contextual methods, and gaining a considerable amount of both theoretical and practical knowledge, it is time to look at the results with regards to the research questions. This section tries to summarize and give a direct answer, complementing the rest of the thesis.

RQ1 *Can contextual data be used to successfully predict fake news in structured data on randomly chosen sources, or are they dependent on good, structured data?*
This thesis has shown that contextual data chosen from strictly structured sources can be used to a high degree of accuracy to predict fake news in a supervised training environment. With results of over 90 % accuracy in predicting fakes using Harmonic Boolean Label Crowdsourcing, and over 85 % accuracy with Logistic Regression, this shows that contextual methods can be used to predict fakes.

Based on the results from [60] together with the results gained from the new dataset gathered in this thesis, it seems that only minor changes in results take place with different sources. The slight differences in results are most likely from the haphazard choice of sources in the new dataset, only taking into account using the same hoax/non-hoax distribution as the original.

It still remains to test the algorithms on unstructured data or on semi-unsupervised data.

RQ2 *Does a web-of-trust augment prediction of fake news, and is it feasible to create one for bigger social networks?*
A web-of-trust has not been created during this thesis. The reason for this is that the calculation complexity when working with over 1 million users was too big to practically gain any reasonable results within the time-span of the thesis.

However, a web-of-trust will most likely be a valuable tool for prediction of different matters, especially in less structured environments, where a pre-generated network can be used to gain insight and map the relationships between entities not directly connected in the data. As it stands now, an all-to-all web-of-trust is not feasible. The need for clustering or otherwise structuring the huge number of nodes is needed. There is also a need of filtering out users that do not contribute to the calculations. When this is done, creating a web-of-trust is not only feasible, but highly sought after.

RQ3 *How volatile are contextual methods compared to each other, and how do their results change based on size, density and others measures?*
This thesis has shown that the algorithms used, Harmonic Boolean Label Crowdsourcing and Logistic Regression, are robust towards changes in size. They give similar results down to 1/16 of the original size when the original density and user/post distribution is maintained. Changes in density and other measures remain to be tested. User density and source dependency are next on the list in testing contextual methods.

7.3 Further work

As mentioned in the previous sections, some improvements will greatly improve the solution further. This section presents ideas that have the potential of greatly increasing the validity and usability of the contextual approach in both language processing and fake news detection.

7.3.1 Preprocessing

The datasets were used directly from the source, in this thesis the Facebook Graph API[16]. Many of the users extracted there did not contain enough information to contribute to the result. The edge amount between them and the rest of the posts and users were just too low. These users should be discarded to test further if they indeed contribute to the results or not, as this has not been tested. Other preprocessing steps that could be taken is to normalize the sources. This does not mirror real-world information, where the majority of the data still are real but will be a good tool in indicating what relationships between users are important for contextual methods.

7.3.2 Unsupervised training

One of the most important experiments that remain doing, is to test the algorithms against bigger datasets which contains unclassified posts and sources. This is important because of the whole range of news articles that exist. They range from research papers to satire and fake news. In addition to the different articles that exist, sources that are not decisively scientific or fake. It is important to find out if the solution is robust enough to classify both types in the same sources. Because news sites gets tricked from time to time, and publish false information by accident. These news are as fake as others but does not normally have the same malicious intent that more specialized sites normally publish.

7.3.3 Complexity minimization

By applying at least some of the preprocessing, the complexity could be sufficiently lowered as to be viable on bigger datasets. The use of simulated annealing or similar best-effort approaches yield good results that can be used as indicators in hybrid solutions and for fact-checking professionals. Finally, both user and post importance must be established. Can the users or the posts be clustered if they are sufficiently equal and whether all the information is needed are questions that needs answering. There is also be situations where adding structured information from other sources like DBpedia will help.

7.3.4 Web-of-trust

By creating a trust network comparable to [61] on top of the already existing solution with the algorithms proposed here, it might be possible to use structured data from one source and use it as a trust network for unstructured data such as news events from newspapers. The trust network will also give an indication of the reliability of sources, and give more data to work with when classifying the information pieces as truthful or false. Achieving this will need bigger datasets than the ones used in this thesis, to solidify the reliability and get statistically significant enough numbers. Determining the weight distribution and impact between users and unstructured sources is paramount and needs to be researched further.

References

[1] Tony Abou-Assaleh, Nick Cercone, Vlado Keselj, and Ray Sweidan. N-gram-based detection of new malicious code. In *Computer Software and Applications Conference, 2004. COMPSAC 2004. Proceedings of the 28th Annual International*, volume 2, pages 41–42. IEEE, 2004.

[2] B Thomas Adler and Luca De Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, pages 261–270. ACM, 2007.

[3] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[4] Benjamin Schiller Andreas Hanselowski, Avinesh PVS and Debanjan Chaudhuri Felix Caspelherr. Athene (ukp lab). `https://medium.com/@andre134679/team-athene-on-the-fake-news-challenge-28a5cf5e017b`.

[5] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[6] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[7] Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. Battling the internet water army: Detection of hidden paid posters. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 116–120. IEEE, 2013.

[8] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM, 2015.

[9] Yimin Chen, Niall J Conroy, and Victoria L Rubin. News in an online world: The need for an “automatic crap detector”. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[10] Sahil Chopra, Saachi Jain, and John Merriman Sholar. Towards automatic identification of fake news: Headline-article stance detection with lstm attention models, 2017.

[11] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[12] Luca de Alfaro, Vassilis Polychronopoulos, and Michael Shavlovsky. Reliable aggregation of boolean crowdsourced tasks. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[13] Maarten De Laat, Vic Lally, Lasse Lipponen, and Robert-Jan Simons. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning*, 2(1):87–103, 2007.

[14] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

[15] Edelman. Where do you get your news? `https://www.edelman.com/p/6-a-m/where-do-you-get-your-news/`.

[16] Facebook. Facebook graph api. `https://developers.facebook.com/docs/graph-api/`.

[17] Factmata. Factmata, building a quality media ecosystem. `http://factmata.com/`.

[18] Faktisk. Faktisk. `https://www.faktisk.no/`.

[19] Weiguo Fan and Michael D Gordon. The power of social media analytics. *Communications of the ACM*, 57(6):74–81, 2014.

[20] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 2016.

[21] FNC. Fake news challenge. `http://www.fakenewschallenge.org/`.

[22] Geeks for Geeks. Reservoir sampling. `http://www.geeksforgeeks.org/reservoir-sampling/`.

[23] Michael Galetzka, Lutz Strüngmann, and Christian Weber. Intelligent predictions: an empirical study of the cortical learning algorithm. *University of Applied Sciences Mannheim*, 2014.

[24] Joshua Gillin. Politifact's guide to fake news websites and what they peddle. `http://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/`.

[25] The Guardian Hannah Ellis-Petersen. Boaty mcboatface wins poll to name polar research vessel. `https://www.theguardian.com/environment/2016/apr/17/boaty-mcboatface-wins-poll-to-name-polar-research-vessel`.

[26] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(7), 2017.

[27] Benjamin D Horne and Sibel Adali. Buzzfeed political news data and random political news data. `https://github.com/BenjaminDHorne/fakenewsdata1`.

[28] http://www.tfidf.com/. What does tf-idf mean? `http://www.tfidf.com/`.

[29] Frank Isca. Google knowledge graph: How it works and its impact on seo. `https://www.weidert.com/whole_brain_marketing_blog/bid/105428/Google-Knowledge-Graph-How-It-Works-And-Its-Impact-on-SEO`.

[30] FORTUNE JOHN PATRICK PULLEN. Fcc and net neutrality: Check to see if your name was used in fake comments. `http://fortune.com/2017/11/29/fcc-and-net-neutrality-check-to-see-if-your-name-was-used-for-fake-comments/`.

[31] Lewis Leong. Fighting fake news: how google, facebook and others are trying to stop it. `http://www.techradar.com/news/fighting-fake-news-how-google-facebook-and-more-are-working-to-stop-it`.

[32] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003*

Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 71–78. Association for Computational Linguistics, 2003.

[33] Pierre Lison. “an introduction to machine learning, 2015.

[34] Marina Litvak and Mark Last. Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*, pages 17–24. Association for Computational Linguistics, 2008.

[35] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[36] Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[37] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[38] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[39] Özlem Özgöbek NTNU. Smartmedia program. `https://www.ntnu.no/wiki/display/smartmedia/SmartMedia+Program`.

[40] Mary Jo Nye. Speaking in tongues. `https://www.chemheritage.org/distillations/magazine/speaking-in-tongues`.

[41] OpenCV. Introduction to support vector machines opencv. `https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html`.

[42] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[43] Neil Patel. The beginner’s guide to google’s knowledge graph. `https://neilpatel.com/blog/the-beginners-guide-to-the-googles-knowledge-graph/`.

[44] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[45] PolitiFact. Politifact. `http://www.politifact.com/`.

[46] Sebastian Raschka. Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*, 2014.

[47] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*, 2017.

[48] Nick Rochlin and Nick Rochlin. Fake news: belief in post-truth. *Library Hi Tech*, 35(3):386–392, 2017.

[49] Yefeng Ruan and Arjan Durresi. A survey of trust management systems for online social communities–trust modeling, trust inference and attacks. *Knowledge-Based Systems*, 106:150–163, 2016.

[50] Victoria L Rubin, Yimin Chen, and Niall J Conroy. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[51] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[52] Sunita Sarawagi et al. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.

[53] Ian Sherr. Facebook ceo zuckerberg discusses battle with fake news. `https://www.cnet.com/news/facebook-fake-news-mark-zuckerberg/`.

[54] Amit Singhal. Introducing the knowledge graph: things, not strings. `https://googleblog.blogspot.no/2012/05/introducing-knowledge-graph-things-not.html?m=1`.

[55] ALEXANDER SMITH and NBC NEWS VLADIMIR BANIC. Fake news: How a partying macedonian teen earns thousands publishing lies. `https://www.nbcnews.com/news/world/fake-news-how-partying-macedonian-teen-earns-thousands-publishing-lies-n692451`.

[56] Statista. Leading social networks used weekly for news in the united states as of february 2017. `https://www.statista.com/statistics/444708/social-networks-used-for-news-usa/`.

[57] Statita. Which of the following types of news is most important to you? https://www.statista.com/statistics/254511/level-of-interest-in-various-news-types-in-the-us/.

[58] Christos Stergiou and Dimitrios Siganos. Neural networks. https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

[59] svms.org. Introduction to support vector machines. http://www.svms.org/introduction.html.

[60] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

[61] Mozhgan Tavakolifard, Kevin C Almeroth, and Jon Atle Gulla. Does social contact matter?: modelling the hidden web of trust underlying twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 981–988. ACM, 2013.

[62] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[63] Wikipedia, the free encyclopedia. Artificial neural network with layer coloring, 2013. [Online; accessed December 28, 2017].

[64] Sean Baird Yuxi Pan, Doug Sibley. Talos in the news. https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html.

APPENDIX A

Acronyms

NLP Natural Language Processing
FNAS Fake News At Scale
NTNU Norwegian University of Science and Technology
ANN Artificial Neural Network
ML Machine Learning
SVM Support Vector Machine
TF-IDF Term Frequency-Inverse Document Frequency
LR Logistic Regression
RHA Regularized Harmonic Algorithm
HBLC Harmonic Boolean Label Crowdsourcing
HTML Hypertext Markup Language
XML Extensible Markup Language
FNC Fake News Challenge
API Application programming interface