CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Investigating Content-based Fake News Detection using Knowledge Graphs

A closer look at the 2016 U.S. Presidential Elections and potential analogies for the Swedish Context

Master's thesis in Computer Science and Engineering

Jurie Germishuys

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2019

Master's thesis 2019

Investigating Content-based Fake News Detection using Knowledge Graphs

A closer look at the 2016 U.S. Presidential Elections and potential analogies for the Swedish Context

Jurie Germishuys

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2019

Investigating Content-based Fake News Detection using Knowledge Graphs
A closer look at the 2016 U.S. Presidential Elections and potential analogies for the Swedish Context
Jurie Germishuys

Supervisor: Richard Johansson, CSE
Advisor: Ather Gattami, RISE
Examiner: Graham Kemp, CSE

Master's Thesis 2019
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2019

Investigating Content-based Fake News Detection using Knowledge Graphs
A closer look at the 2016 U.S. Presidential Elections and potential analogies for the Swedish Context
Jurie Germishuys
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

In recent years, fake news has become a pervasive reality of global news consumption. While research on fake news detection is ongoing, smaller languages such as Swedish are often left exposed by an under-representation in research. The biggest challenge lies in detecting news that is continuously shape-shifting to look just like the real thing — powered by increasingly complex generative algorithms such as GPT-2. Fact-checking may have a much larger role to play in the future. To that end, this project considers knowledge graph embedding models that are trained on news articles from the 2016 U.S. Presidential Elections. In this project, we show that incomplete knowledge graphs created from only a small set of news articles can detect fake news with an F-score of 0.74 for previously seen entities and relations. We also show that the model trained on English language data provides some useful insights for labelling Swedish-language news articles of the same event domain and same time horizon.

Keywords: fake news, knowledge graphs, embedding models, natural language processing, generative models, Swedish.

Acknowledgements

I would like to thank my academic supervisor Richard Johansson for his continuous and invaluable support during the project. I would also like to thank my industrial supervisor, Ather Gattami for his creative insights, brainstorming sessions and contributions to the thesis as well as providing the opportunity to pursue this thesis at RISE.

Jurie Germishuys, Gothenburg, August 2019

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
API	Application Programming Interface
AUC	Area Under Curve
AUPRC	Area Under Precision Recall Curve
GAN	Generative Adversarial Network
GNMT	Google Neural Machine Translation
LCWA	Local Closed-World Assumption
LSTM	Long-short Term Memory Network
NER	Named Entity Recognition
NLP	Natural Language Processing
PCA	Principal Components Analysis
POS	Part of Speech
RDF	Resource Description Framework
RISE	Research Insitute of Sweden
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
U.S.	United States of America

1

Introduction

1.1 Context

Today, malevolent parties use false narratives to influence opinions around the world. Brought to light during the Trump campaign in 2016, the term ”fake news” is now a globally relevant problem. Unfortunately, the pace of fake news development is fast approaching a point at which the average human will be unable to distinguish fact from disguised fiction. Some advanced generative models such as GPT-2 released by OpenAI have already reached a point at which the creators considered the model content to be too ”human-like” for public release, prompting fear and caution about the acceleration of nearly undetectable artificial content [22].

There is also now a strong financial incentive for content style that is constantly evolving and escapes detection by state of the art fake news detection algorithms. In small villages as remote as Veles in Macedonia, many now see fake news generation as a lucrative career path that even has an official curriculum and university to entice a growing number of young people to generate fabricated articles [10]. These global trends necessitate the exploration of updating, temporal models capable of handling streaming data and complex patterns [30].

It is predicted that by 2022, the developed world will see more fake news than real information on the internet [21]. New techniques in artificial intelligence are leading the charge in the production of such fakes, but equally offer us the opportunity to analyse huge amounts of data and verify content to combat the influx of misinformation [16].

Sweden has, of late, also seen a prime example of the pervasiveness of fake news. In the recent 2018 election, an Oxford study found that 1 in 3 articles shared on social media during the election period were indeed false [14]. It is clear that there is a need for smaller language communities to be able to assess the veracity of their news sources and their associated claims. This project aims to form part of a larger body of research undertaken by the Research Institute of Sweden (RISE) to develop a workable model for the Swedish context.

1.2 Challenges

The detection of fake news presents a slew of challenges, some of which are discussed further in this report, including:

1. Fake news is difficult to define concisely and consistently, as its nature changes significantly over time. This means that while in the past, purely stylistic approaches were quite successful, the convergence of fake news to the writing style of real news will likely lead to degrading performance. Understanding how fake news is generated could, therefore, lead to insights that are pro-active rather than retrospective in fake news detection.

2. Recent studies have shown that fake news stories spread more quickly than they can be identified, so the sources of fake news also need to be detected rather than focusing only on individual articles. [26]

3. Ground-truth verification of article claims in aggregate is not always possible since studies have shown that humans are ”average” at detecting deception, with an accuracy in the range 55-58% [24].

4. Natural language processing approaches are susceptible to adversarial attacks (e.g. a fake news article produced by a GAN algorithm) that mimicks the look and feel of a trusted news source [38].

5. 'Fake news' is a heavily context-dependent and time-dependent classification, as news is only current for a certain period of time, and corrections or retractions are common.

6. The topic of fake news detection in languages other than English has been underrepresented in research and thus supervised approaches that work well in English do not perform as well in non-English domains. One of the main reasons for this is the lack of labeled training data.

1.3 Contributions

The primary contribution of the project is the design of a lightweight fact-checking model, which is centered around key controversial events or topics in a well-defined time-window. This project also focuses on model explainability, as the system proposed in the project aims for a human in the loop design, meaning that the model should augment the ability of the end-user rather than be a black-box automation solution. As far as the author knows, this is the first attempt to use a knowledge graph approach for fake news detection in Swedish. It is the hope that this project will provide a baseline dataset for continued research into fake news detection in Swedish and other smaller languages.

1.4 Goals

The project aims to develop a knowledge graph embedding model, given a context of prior knowledge, and evaluate the following end-goals:

1. Given a statement, score the statement based on knowledge graph embeddings.

2. Given a series of statements (in the form of an article), aggregate the output of the knowledge embedding model for a single statement to determine whether an article is most likely to be real or fake. Each statement will be traversed as a triple (h, t, r), where h refers to a head entity, t to a tail entity and r to a relational vector in the knowledge graph.

3. Create a reference Swedish dataset with labels for use in future fake news detection research using a model based on English language data.

1.5 Scope

This research project focuses on only a limited set of news articles over a given event horizon within a given time period. It is thus not designed to represent a large body of knowledge, but rather a focused set of articles that represent ”fake news” within a particular context over a particular time period.

1.6 Thesis Structure

The thesis is structured in the following way. In Chapter 2, Related Work, the theoretical foundations of knowledge graphs as well as the problem of 'Fake News' are discussed. In Chapter 3, the methodology of the chosen embedding models is explored. The results from these methods are collected in Chapter 4, including an evaluation of outcomes and a discussion of their limitations. The final chapter contains an answer to the goals set in Chapter 1 as well as a dicussion on future work in this area.

2

Related Work

2.1 Defining Fake News

There is no universally accepted definition of 'fake news' since many types of news or information could qualify under the view that fake news is simply spreading misinformation or rumors [8]. However, some definitions more explicitly state a need for fake news to also have malicious intent [3]. Often, social media has been the medium for the propagation of such news, but this channel will not be considered here since we do not consider social media websites to be news sources but rather news aggregators.

The nature of fake news has not been static over time, constantly morphing in parallel with attempts to detect it — therein lies the most difficult part in defining 'fake news' for more than a limited window of time. Researchers at the University of Washington recently released a generative model called 'GROVER' which claims to be able to distinguish fake news generated by a neural network from human-written fake news with an accuracy of 92% [36]. Grover generates an article based on a particular title and author, e.g. the title "Trump Impeached" generated the following fake article:

> *The U.S. House of Representatives voted Wednesday on whether to begin impeachment proceedings against President Donald Trump, seeking to assert congressional authority against the president just days after the release of special counsel Robert Mueller's final report on Russian interference in the 2016 election.*

The definition used in this research project, considers fake news to be *unverified news information purported as fact from a given news outlet over a pre-defined time horizon on a particular event domain.* This definition applies regardless of the origin or intent on the author's part. In this setting, the intent of the news article falls outside the scope of a content-based model as any statement will be taken at face value. This allows for a sufficiently broad definition for the classification task as

set out in Chapter 1. It also aligns broadly with the definition of ”false news” as outlined in Zhou and Zafarani.

The following is an example of fake news article that clearly stands out stylistically, it uses hyperbolic, subjective language to describe the parties involved that put forward a particular point of view:

> *’BOOM! CHARLIE DANIELS Nails Obama And Democrats In Just One Tweet. Obama has been low key in the past few months even as he campaigned for a losing Hillary Clinton.Suddenly Obama and the Democrats decided Obama and the Democrats CARE about Russia and so Obama got all tough with Putin, which is sorta hilarious if you think about it. Dump on top of that the mess in Israel, Obamacare, the Iran fail, millions of Americans out of work, and the attempts at forcing states to fund Planned Parenthood, and you have a nice big MESS that Trump and Trump administration will have to figure out.*

However, in an ever-increasing number of cases, the language is not the main discriminator [22]. In the following case, we see fairly objective language that simply describes a sequence of events as though it were factual, and instead leaves the reader to follow the author’s logic and to draw conclusions based on this sequence. These news articles are the primary candidates of the models presented in this research project.

> *’EXCLUSIVE: Ex-Bernie Delegate Reveals Why Ex-Bernie Delegate Fled Democratic Party for the Greens. Roving political analyst Stuart J Hooper drops in the see what was happening as Bernie Sanders hit the western college campuses on to campaign for Hillary Clinton. The following is an interview with an ex-Bernie delegate who, following the DNC collusion with the Hillary Clinton camp to kill the Sanders campaign, has since left the Democratic Party to support Dr Jill Stein and the Green Party. Ex-Bernie Delegate explains how Sanders was coerced into backing the Hillary Clinton campaign.’*

2.2 Fake News detection

Fake news detection approaches can be loosely divided into three main categories: content-based, style-based and propagation-based.

2.2.1 Content-based models

Content-based or knowledge-based approaches, also known as ”fact-checking”, involve using a ground-truth knowledge base, usually populated by experts or crowd-sourced, in order to compare the information from one source to a trusted or verified source. This can be done both manually or automatically. One manual approach is to use human experts (usually journalists or political scientists) to score statements. This is used by the fact-checking website Politifact, which scores statements by prominent political figures in the United States, and has also developed a scorecard for news articles surrounding political events, such as the 2016 U.S. presidential election. With the large amounts of information available today, automatic approaches using knowledge bases have increased in popularity as the need for scalability and speed of retrieval becomes increasingly important. These knowledge bases are constructed by first extracting facts from the open web, and then processing this raw data into Resource Description Framework (RDF) triples, known as Automatic Open Information Extraction [37].

In an ideal setting, having access to perfect information would allow these facts to be easily corroborated or refuted. However, even in the case of automatic knowledge extraction, knowledge bases are unable to keep up with the current pace of streaming news information. They also tend to be sparse, which means that links between parts in disparate areas of the graph cannot easily be made. In addition, a large amount of knowledge base information is not useful in fake news detection, as mostly more contentious and less axiomatic information will be presented. For example ”Immigrants are a net drag on the economy” is a compound statement which is not in itself true or false, but puts forward a more complex argument that first need to be broken down into individual assertions that can be verified. This leads us to explore models that are able to learn the links between different entities and relations given a knowledge base, and which can be used for sparser or more incomplete knowledge graphs.

2.2.1.1 Knowledge Embedding Models

Knowledge graphs are data structures that represent knowledge in various domains as triples of the form (h, t, r), where h refers to the head entity, t to the tail entity and r to the relation between them. An example thereof is (Stockholm, isCapitalCity, Sweden). Knowledge graphs are a popular tool to represent the information inside knowledge bases, which is essentially a technology used to store various forms of information. They have also become a popular tool used in machine learning and artificial intelligence (AI), as the graph structure allows more complex relations between entities to be exploited, particularly in the domain of natural language processing. Popular applications include question-and-answer (QA) systems for voice assistants, parole decisions, credit decisions, anomaly detection and fraud detection [27].

A knowledge graph embedding approach converts the entities and relations from a knowledge graph into low-dimensional vectors, which are more suitable for use in machine learning algorithms. These models are particularly appealing because they are transparent and explainable, since model decisions can ultimately be traced back to paths in the knowledge graph. One such model uses existing open knowledge bases in English such as DBpedia, which showed that even incomplete knowledge graphs could provide useful results for fake news detection by evaluating statements using an existing context of facts (i.e. fact-checking). Additionally, this model demonstrated that fake news detection was possible with F-scores around 0.8 using only news articles and no ground-truth knowledge base [20]. This paper forms the primary theoretical basis for the research questions in this project.

Knowledge embedding models are not new, but the application of knowledge graphs to fake news detection is a relatively novel idea. Knowledge embedding attempts to bridge the gap between graph-structured knowledge representations and machine learning models. In a related domain, spam classification optimisation has made use of knowledge graph embeddings as an input to the deep network that determines whether a particular review text was written by a particular author, as a way of solving the so-called "cold-start problem" in spam classification, which refers to the fact that it is difficult for the model to classify a new review from an unknown source as "spam" or "not-spam" [29].

2.2.2 Style-based models

Style-based approaches focus on the way in which fake news articles are written. This includes the use of language, symbols and overall structure. These methods are based on the core assumption that the distribution of words and expressions in fake news is significantly different from real news [37].

In essence, a new article can then be classified as 'fake' or 'true' based on a feature set which is either crafted manually according to rules (e.g. the number of exclamation points) or extracted automatically (e.g. through a deep learning model). Often these approaches involve machine learning algorithms that are able to extract structure-based as well as attribute-based features, such as the word count, use of hyperbole and sentiment.

Earlier papers on fake news identification used TF-IDF (term frequency inverse document frequency) to encode the headline and the body of a news article separately, known as stance detection [23]. This involves developing a probabilistic model of the language used in fake news articles by counting the number of times a particular word appears in a range of documents and then dividing that by the number of documents in which the word appears. After encoding, they were compared using a single-layer neural network and computing the softmax over the following categories: "Agree", "Disagree", "Discuss" and "Unrelated". If there was disagreement between the headline and the article body, the article was more likely to be classified

as ”fake”, and vice-versa. The largest competition held on fake news detection in 2017 focused on this approach, where a team combining a deep neural network and an ensemble of decision trees won with an accuracy of 82% in stance detection.

Other studies have focused on the style of the URL and attributes linked to the source rather than on the content of the article itself. Using features such as the content of a news source's Wikipedia page and information about the web traffic it attracts, the classifier was able to attain an accuracy of around 65% [5].

The results above illustrate the difficulty in pinning down the stylistic nuances in fake news, with detection rates well below the level required to make these detectors effective. Based on the results from the paper by Baly et al., MIT recently claimed that even the best detection systems were still ”terrible” at identifying fake news sources [13]. Thus, the detection of false news in news articles based on stylistic features alone requires deeper investigations into less overt patterns, supported by theories from closely-tied domains, such as journalism [37].

2.2.3 Propagation-based models

Another approach has emerged recently, focusing rather on the propagation of news on social media as a measure of its veracity. These approaches have focused on studies showing that fake news spreads faster than and about 100 times further than true news in the domain of politics [26]. One measure of this spread is a cascade, which is a network structure illustrating how a news article moves from the original poster to how it is shared by other users, usually in a social media setting. Another measure looks at the stance taken by users to a news post, which translates to computing the distance between the user posts in what is termed a ”stance network”. If there is a large degree of disagreement, it points to an increased likelihood of fake news [37].

2.3 Fake News in Swedish

The lack of research into fake news for smaller languages risks exposing readers to unprecedented amounts of unfiltered and unverified information. An Oxford Internet Institute study found that the proportion of fake news shared on social media during an election was the 2nd highest during the 2018 Swedish election, the first being the 2016 presidential elections in the United States. It also far outpaced other European countries, underscoring the importance of this issue in the Swedish context. Additionally, in contrast to the United States, the fake news problem was much more likely to be homegrown rather than externally-produced, with only around 1 percent of fake content traced back to foreign sources [14]. This situation calls for approaches that use smaller amounts of data that attain classification results similar to those in the most spoken languages, such as English and Mandarin.

3

Methods

This chapter starts by defining two important knowledge embedding models, TransE and B-TransE, and their training procedures. Then, the application of these models to the fake news classification task is explored. Other important methodological considerations, including the choice of datasets and processing of triples are also dealt with. The final part of the chapter elaborates on the transition from English language data to Swedish language data and finally highlights the evaluation metrics used to score the various model implementations.

3.1 TransE

The simplest form of knowledge graph embedding model is based on mapping the translation of an entity to another via a relation vector, r. The goal of TransE is to embed entities and relations into low-dimensional vectors. The embedding returns this vector as a tuple of vectors (h, t, r), where h corresponds to the embedding vector of the head (subject), r the embedding vector of the relation and t the embedding vector of the tail (object). The idea here is that $h + r \approx t$ if $(h, t, r) \in T(h, r)$, i.e. that the relation is a translation of the entity vector [7]. This is illustrated clearly in Figure 3.1.

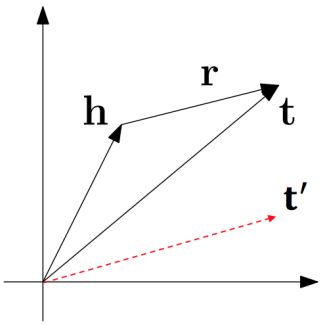


Figure 3.1: Plot showing the embedded vectors, h, t, r. It is clear that the triple (h, t, r) represents a triple from the embedded knowledge graph, whereas (h, t', r) is not likely to be a triple found in this embedded knowledge graph.

3.1.1 TransE training

Each low-dimensional relation and entity embedding vector is randomly initialised by sampling from a uniform distribution. At the start of each iteration, each of these is then normalised. The algorithm then samples a small batch of statements from the training set. Then, for each statement (triple) in the batch, the algorithm constructs a negative corrupted triple (either by corrupting the entities, relations or both). The batch is then enlarged by adding the corrupted triples. The randomised embeddings are then updated by minimising the model objective (loss function, $\mathcal{L}$) using gradient descent, which in this case is the gradient of sum of the margin, γ and the difference between some dissimilarity measure, $d(\mathbf{h} + \mathbf{l}, \mathbf{t})$, which measures the error between a corrupted triple and a correct triple. The algorithm stops based on the prediction performance on a validation set of triples. The main idea here is that the model should learn to distinguish between corrupted and correct triples from the knowledge graph. The pseudo-code for this algorithm is presented in Figure 3.2.

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
  $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$
  $\mathbf{e} \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$
loop
  $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$
  $S_{batch} \leftarrow$sample(S, b) // sample a minibatch of size b
  $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
  for $(h, \ell, t) \in S_{batch}$ **do**
    $(h', \ell, t') \leftarrow$sample$(S'_{(h,\ell,t)})$ // sample a corrupted triplet
    $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
  end for
  Update embeddings w.r.t. $\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla \left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$
end loop

Figure 3.2: Pseudo-code for the implementation of the training algorithm for TransE. $\mathcal{L}_+ = max(\mathcal{L}, 0)$

3.2 Problem Formulation: News Article Classification

Assume a news article is represented by a set of statements in the form of RDF triples (h_i, t_i, r_i), i = 1, 2,. . . , n. Let K_T refer to a knowledge graph containing a set of labelled true news articles denoted as A_{t_j}, j = 1, 2,. . . , m. Let K_F refer to a knowledge graph containing a set of labelled fake news articles denoted as A_{f_j}, j = 1, 2,. . . , m.

The task of evaluating the authenticity of each news article A_j is to identify a function S that assigns an authenticity value $S_i \in \{0, 1\}$ to A_j in where $S_i = 1$ indicates the article is fake and $S_i = 0$ indicates it is true [37].

Since we are dealing with small datasets that could not possibly encapsulate all 'true' or 'fake' knowledge, we have to make an assumption about how we deal with unseen triples.

Local Closed-world Assumption [11]:

The authenticity of a non-existing triple is based on the following rule: suppose $T(h, r)$ is the set of existing triples in the knowledge graph for a given subject h and predicate r. For any $(h, t, r) \in T(h, r)$, if $|T(h, r)| > 0$, we say the triple is valid for evaluation; if $|T(h, r)| = 0$, the authenticity of triple (h, t, r) is unknown.

The Local Closed-world Assumption means that triples that involve entities and relations not yet seen by the model are discarded during the evaluation phase.

3.3 Single TransE models for fake news detection (Pan et al.)

The training procedure for fake news detection is largely the same as the standard procedure for TransE. For fake news detection, the TransE model is trained either exclusively on fake news articles to construct K_F or true news articles to construct K_T from the training set.

The dissimilarity measure or bias for a particular triple (h, t, r), $d(\mathbf{h} + \mathbf{r}, \mathbf{t})$, which is computed as in Eq (3.1), uses the L2-norm as a distance metric.

$$d_b\left(\text{triple}_i\right) = \|\mathbf{h}_i + \mathbf{r}_i - \mathbf{t}_i\|_2^2 \tag{3.1}$$

In the case of the single TransE model, classification of a news article is performed by aggregating the computed biases for each statement in a news article, (h_i, t_i, r_i), i = 1, 2,..., n. The aggregation can be done either as the average bias as in Eq (3.2) or the maximum bias across triples as in Eq (3.3) in the article.

$$d_{avgB}(TS) = \frac{\sum_{i=1}^{n} d_b\left(\text{triple}_i\right)}{|TS|} \tag{3.2}$$

where $|TS|$ refers to the size of the test set.

$$d_{\max B}(TS) = \underset{i}{\operatorname{argmax}}\, d_b\left(\text{triple}_i\right) \tag{3.3}$$

where $\text{argmax}_i\, d_b\,(\text{triple}_i)$ refers to the triple with maximum bias for each article in the test set.

The aggregated bias is then compared to a relation-specific threshold, r_{th}, which is computed as the threshold that maximises the accuracy at the article level on the validation set.

Example:

Assume we are working with the knowledge graph created from the True news articles, K_T. Say (Löfven, supports, peace) produces ([1.0,1.5,1.6], [1.0, 2.0, 1.7], [2.0, 3.5, 3.3]) in our embedding model. When we have a new triple from an article, say (Löfven, supports, demilitarization), this produces the tuple ([1.0,1.5,1.6], [1.0, 2.0, 1.7], [2.0, 3.5, 4.0]) from our embedding model, assuming $d = 3$. The magnitude of the bias is then calculated as the norm of ([1.0,1.5,1.6]+[1.0, 2.0, 1.7]) – [2.0, 3.5, 4.0] = [0, 0, 0, -0.7], which is 0.7. With a relation-specific threshold for (*support*) of 1.5, we can say that in the low dimensional space, these vectors are likely to lie close each other, and therefore it is unlikely to be fake news. The reverse is true if the bias is high.

3.3.1 B-TransE model for fake news detection

A single TransE model trained on a small amount of data has two main sources of error. This occurs because an article could have a large dissimilarity or bias in both the single 'True' and 'Fake' TransE models, resulting in a contradicting and inconclusive outcome. To overcome this, Pan et al. propose a novel approach that compares the dissimilarity functions of both models. The model with the lowest dissimilarity score is then chosen as most likely to represent that particular article. The dissimilarity measures for the B-TransE model are calculated as shown in Eq (3.4).

$$\begin{aligned} d_b\left(\text{triple}_{t_i}\right) &= \|\mathbf{h}_{t_i} + \mathbf{r}_{t_i} - \mathbf{t}_{t_i}\|_2^2 \\ d_b\left(\text{triple}_{f_i}\right) &= \|\mathbf{h}_{f_i} + \mathbf{r}_{f_i} - \mathbf{t}_{f_i}\|_2^2 \end{aligned} \tag{3.4}$$

In the B-TransE model, the aggegrated bias for each article is compared and the model with the lowest bias is selected as the prediction. Once again, aggregation can be done as an average or as a maximum bias across triples for each article, as shown in Eq (3.5) and Eq (3.6) respectively.

$$\begin{aligned} d_{mc}(N) =& 0, if \operatorname*{argmax}_i d_b\left(\text{triple}_{f_i}\right) < \operatorname*{argmax}_i d_b\left(\text{triple}_{t_i}\right) \\ & d_{mc}(N) = 1, \text{ otherwise} \end{aligned} \tag{3.5}$$

$$\begin{aligned} d_{ac}(N) =& 0, if \operatorname*{argmax}_i d_{avgB}\left(\text{triple}_{f_i}\right) < \operatorname*{argmax}_i d_{avgB}\left(\text{triple}_{t_i}\right) \\ & d_{mc}(N) = 1, \text{ otherwise} \end{aligned} \tag{3.6}$$

Based on the findings of Pan et al. as well as the author's own investigation, the max bias aggregation method was chosen for this project.

3.3.2 Hyperparameters

Table (3.1) lists the important hyperparameters in the OpenKE implementation of TransE [12]. The optimal hyperparameters were informed by experiments by Krompaß et al. [18] It should be noted that for the purposes of this project, the validation and the training set are the same, as we have not done any hyperparameter tuning, and overfitting is not a concern since we want the model to memorise as many of the facts presented as possible.

Parameter	**Description**	**Value**
T	Training times	5000
α	Learning rate	0.001
γ	Margin	2
k	Embedding Dimension	50
s_e	Entity negative sampling rate	10
s_r	Relation negative sampling rate	0

Table 3.1: Optimal configuration parameters. Training is done using the Adam optimizer and early stopping with a patience of 20 and ϵ of 0.01.

3.4 Datasets

English: The ISOT fake news dataset from the University of Victoria contains around 40,000 articles collected from various global news sources and labelled either "true" or "fake" according to Politifact. The articles labelled as "fake" were categorised as "unreliable" by Politifact [1, 2]. It should be noted that the "True" articles and "Fake" articles have different types or subjects. For "fake" news, these groups are "Government News", "Middle-East", "US-News", "Left-News", "Politics" and "News". For the "true" news, these groups are "World-News" and "Politics-News".

This dataset was chosen as it was not possible to obtain the news article dataset used by Pan et al. for replication.

Swedish: Dataset sourced from news articles provided by Webhose.io. The dataset includes a collection of 234,196 articles crawled from 133 news sources during October 2016. This dataset is unlabelled. [31]

3.5 Data preprocessing

The original dataset was filtered in order to reduce redundancies and to focus the articles on the event domain to improve generalisation. This was motivated by Vosoughi et al.'s [26] finding that bursts of fake news are often initiated by important events (e.g., a presidential election). They are therefore context-sensitive and time-limited in nature.

Thus, in the first instance, to replicate the results of Pan et al., it was necessary to limit the domain to the U.S. elections of 2016. For the analysis, a dataset of fake and true articles was created using the following pipeline:

True articles:

1. Choose only the subset with the category 'Politics-News'
2. Select a subset of the data containing articles published between 2016-08 and 2016-11 (months immediately before and after the election date)
3. Lastly, select a subset of the data containing articles with the keywords "election", "trump", "hillary" and "obama"

Fake articles:

1. Choose only the subset with the categories 'politics', 'US News', 'Government News'
2. Select a subset of the data containing articles published between 2016-08 and 2016-11 (months immediately before and after the election date)
3. Lastly, select a subset of the data containing articles with the keywords "election", "trump", "hillary" and "obama"

After selecting these subsets the dataset, 1428 fake articles and 1463 true articles remain. Then each article is summarised by extracting the title and the two first sentences of the news article. These summaries are then used to generate triples. This was done in order to reduce redundancies and decrease model training times.

3.5.1 Triple extraction

The Python wrapper package of Stanford OpenIE is used to extract triples in the form (h, t, r) from each summarised news article. The Stanford OpenIE package extracts binary relations from free text. The first step in this process is to produce a set of standalone partitions from a long sentence. The objective is to produce "a set of clauses which can stand on their own syntactically and semantically, and are

entailed by the original sentence". This process is informed by a parsed dependency tree and trained using a distantly supervised approach. It is supervised in the sense that it creates a corpus of noisy sentences that are linked via a known relation (i.e. subject, object pairs). This is then used for distant supervision to determine which sequence uses the correct relation, i.e. which subject and object return the known relation [4]. This process is illustrated in Figure (3.3).

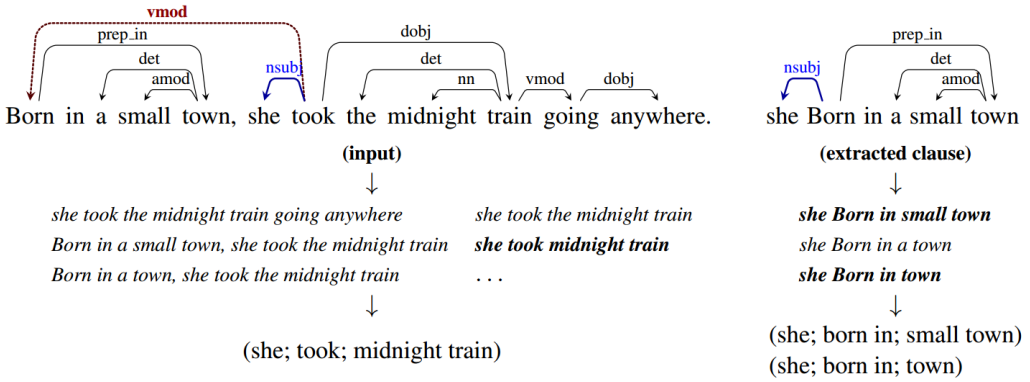


Figure 3.3: An illustration of the approach by Angeli et al. [4] is used to build a triple from an extract of the lyrics from the hit song 'Don't stop believin' from Journey.

The Stanford OpenIE does not provide perfect extractions, and in around 18% of the available news article summaries, the extractor did not provide any triples at all. In particular, we noticed that the model does not deal with negated verbs and sentences containing multiple verbs. An example of this is the sentence "Paul Ryan tells us that he does not care about struggling families living in blue states".

On the other hand, the Stanford OpenIE model also has a pronounced side-effect of over-generation, which means that if multiple verbs are present, it will generate multiple possible triples for a single sentence, which creates a large amount of noise in the triples for each article, with many near-duplicate triples. Figure 3.4 shows an example of how OpenIE breaks down a complex statement from our corpus.

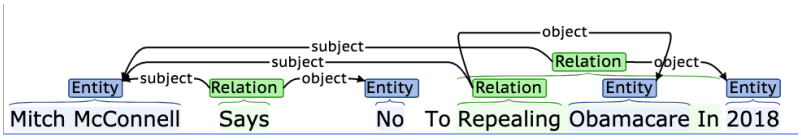


Figure 3.4: An example of overgeneration by Stanford OpenIE by Angeli et al. We can see that three triples are generated from this sentence, with none of them including the full relation "says no to" or the full tail entity "no to repealing Obamacare in 2018"

From the 2891 articles available, triples were successfully extracted from 2373 articles (1038 fake, 1335 true). 1000 articles were then randomly sampled from each category to equalise their representation. The dataset was then split into a training set of 800 and a test set of 200 for each classification. Table 3.2 shows the number of entities, relations and triples extracted from the training set of 2000 articles.

Description	Count
Entities	3K
Relations	6K
Triples	10K

Table 3.2: Training dataset statistics for the 2016 U.S. presidential election data.

3.5.2 Triple processing

The triples are processed according to a pipeline to reduce the noise found in the entities and relations extracted from the training set news articles.

Co-reference resolution:

Co-reference resolution refers, amongst other things, to the process of disambiguating pronouns. For e.g. "James bought cheese. He found it to be tasty.", would be converted to "James bought cheese. James found it to be tasty." This is done using the NeuralCoref package [32].

Relation simplification and lemmatization:

Figure 3.4 shows the distribution of relations after relation simplification and lemmatization. Firstly, the main verbs are extracted from the relations using the Spacy POS tagger. The lemmatization process then converts each word into a normalised form. In this case, each relation is transformed into its infinitive form using the NLTK Lemmatizer. E.g. "are" to "be" and "has" to "have". This deals with verbs that have the same meaning but are expressed in different forms and thus helps to reduce the noise in our relations.

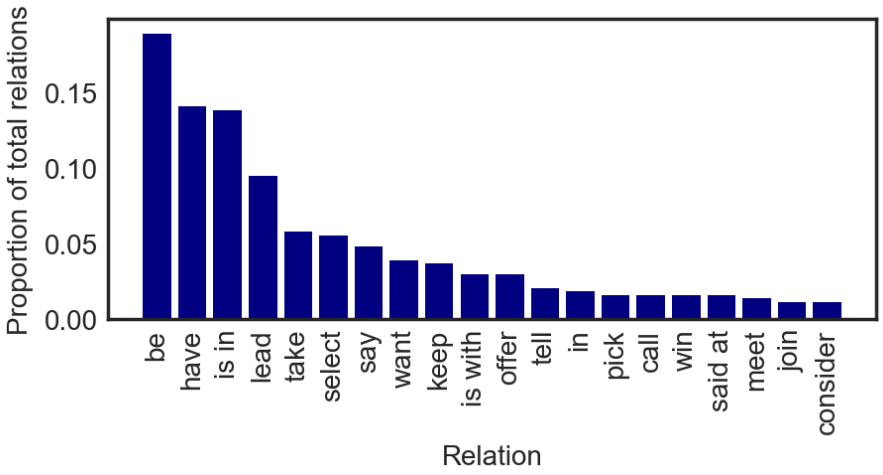
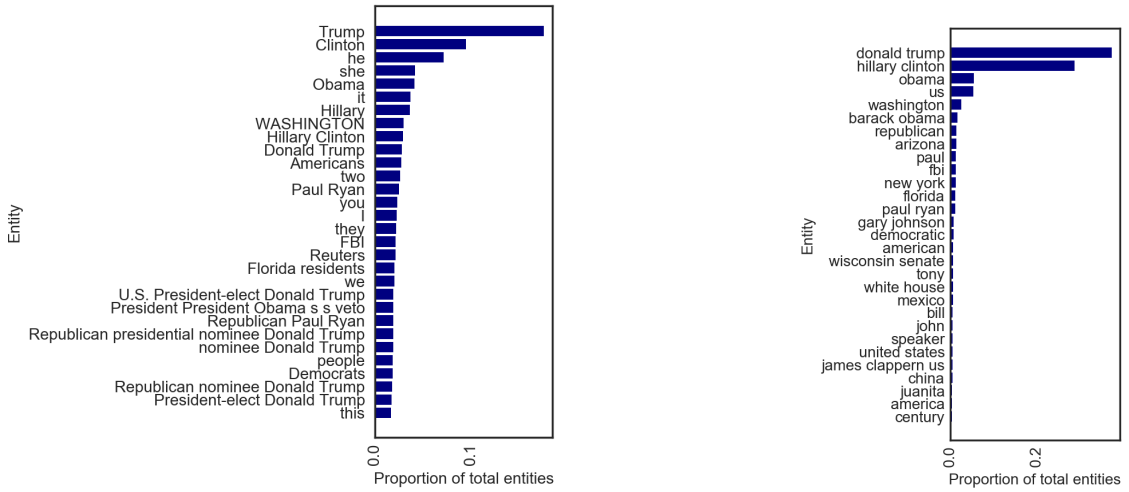


Figure 3.5: Histogram showing the long-tail nature of the relations in the training set. The top 4 relations cover 56.2% of the total relations.

Entity simplification and alignment:

The SpaCy Named Entity recognizer is trained on news, blogs and comments. Using the pre-trained 'en_core_web_lg' model, the pipeline of POS tagging, followed by parsing and then named entity recognition is used. The named entity recognizer module is used to extract persons, locations and organizations and other recognisable entities from longer entities to improve on the long-tail distribution of entities (as shown in Figure 3.5). Stopwords such as "is", "it" etc. are also removed using NLTK, as these are likely to be uninformative entities. The same pipeline is applied to head and tail entities [1].



(a) Entity histogram before pre-processing

(b) Entity histogram after pre-processing

Figure 3.6: Distribution of entities before preprocessing in (a) shows the redundancies of longer entities and uninformative entities such as "he" or "she". The pre-processing pipeline refines the concentration of entities in (b), particularly through the use of co-reference resolution and entity alignment with named entities.

3.6 Extension of Stanford OpenIE and TransE to Swedish

The TransE model is language-agnostic, however the Stanford OpenIE wrapper currently only supports automatic information extraction in English. Approaches to enable automatic information extraction in languages other than English have focused on rebuilding the NLP pipeline. This involves creating a bespoke POS tagger, a language-specific dependency parser, a NER model and training a distantly supervised model to replicate the Stanford OpenIE model. Alternatively, rule-based approaches have also been used to fill the final gap, which so far has only been attempted in German and Chinese [6].

[1] https://spacy.io

Although the first two building blocks (POS tagger and a dependency parser) have been designed by Swedish researchers, an open-source automatic information extractor does not yet exist for Swedish [17]. It was seen as outside of the scope of this project to create such an extractor. Instead, the approach used in this project relies on a proposed transfer learning approach, which attempts to map the embeddings from the 2016 U.S. Presidential Election data over to the Swedish context, using the same pre-specified parameters, i.e. referring to the same entities and written over the same time period (2016-07 to 2016-12). This approach creates labels for the unlabelled Swedish data using the embeddings trained from the English language model.

3.6.1 Data Preprocessing

The pipeline used for the 2016 U.S. Presidential Election data was also applied to the Swedish news dataset. The size of the Swedish news dataset after pre-processing was 1441 articles. The results of the preprocessing is shown in table (3.3).

Description	**Count**
Entities	4K
Relations	8K
Triples	20K

Table 3.3: Training dataset statistics for the Swedish news dataset.

3.6.2 Translation

The summaries generated for Swedish news articles are translated using the Google Cloud Translation API. This cloud-based service connects directly to Google's Neural Machine Translation model (GNMT). The GNMT model uses a technique called 'Zero-Shot Translation' to bypass the need to store the same information in many different languages (e.g. in a knowledge base), and instead is trained to understand the correlation between different languages [15].

3.6.3 Labelling

The Swedish news articles are initially unlabelled. The news articles are then assigned a label of "true" or "false" based on the B-TransE model trained on all of the English articles in the specified domain and time period.

3.7 Evaluation metrics

In order to evaluate our binary classification model and compare between models, a confusion matrix is used.

	Predicted Positive	Predicted Negative
Labeled Positive	True positive (TP): Articles correctly classified as fake	False negative (FN): Articles incorrectly classified as true
Labeled Negative	False positive (FP): Articles incorrectly classified as fake	True negative (TN): Articles correctly classified as true

Table 3.4: Confusion matrix with explanation of outcomes

From the confusion matrix in Table (3.4), a collection of performance measurements can be calculated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.10}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.11}$$

Accuracy Eq. (3.7), shows the overall performance of the classifier, given a symmetric data set (equal distribution of positive and negative examples).

Precision Eq. (3.8) gives an indication of the proportion of those articles predicted as fake, which were in fact fake.

Recall Eq. (3.9) gives an indication of the proportion of those articles which were in fact fake which were accurately predicted as such. This is often referred to as the sensitivity of the classifier.

Specificity Eq. (3.10) gives an indication of the proportion of actual true articles that are predicted as true.

Finally, the F-score Eq. (3.11) is a measurement of the balance or harmonic mean between precision and recall. This is often used as an overall performance metric of the classifier [25].

3.7.1 Precision recall curve

Precision-recall curves plot the relationship between precision and recall (or sensitivity). This curve focuses on the model's ability to identify all the fake news articles, even if this translates into a higher number of FP. A useful summary statistic from the precision-recall curve is the AUPRC, which quantifies the ability of the model to detect fake news articles. This can be thought of as an expectation of the proportion of fake news articles given a particular threshold, and is shown in Eq (3.11). An AUPRC output equal to the proportion of true positives would correspond to a random classifier.

It has also been shown that when detecting rare occurrences (as is the case with fake news), the area under precision-recall curve (AUPRC) metric is preferable to the conventional area under curve (AUC) metric, which is the area under the Recall vs Specificity curve, as it better summarises the predictive performance of the classifier.

$$\text{AUPRC} = \sum_n (R_n - R_{n-1})\, P_n \tag{3.12}$$

where P_n and R_n are the precision and recall at the nth threshold [19].

4

Results

This chapter first explores the predictions from the TransE and B-TransE models in the fake news classification task on the 2016 U.S. Presidential election data. A few key insights from the investigation are presented to better understand the outcomes from the model. Then, the focus shifts to the Swedish context and how these insights could be translated. The chapter concludes by examining some of the important limitations of the aforementioned models.

4.1 Fake News Classification

The results of both the single TransE model and the B-TransE model are shown in Table 4.1. From this table, we can conclude that the TransE (Fake) model performs extremely poorly in terms of recall with a higher number of false negatives. This is somewhat expected as its main source of error is the classification of high bias triples as 'true'. However, it manages to avoid a large number of false positives and has a high precision at 0.82. Nonetheless, the single TransE model produces a very poor classifier in terms of F-score.

Model	**Precision**	**Recall**	**F-Score**
TransE (Pan et al.)	0.75	0.78	0.77
TransE (Fake)	0.82	0.19	0.30
B-TransE (Pan et al.)	0.75	0.79	0.77
B-TransE (English)	0.68	0.67	0.67

Table 4.1: 5-fold cross-validation results from the evaluation of the full test set.

The table also highlights that the B-TransE model performs the best overall, improving on nearly all metrics, with an improvement of 37 percentage points in absolute terms over the single TransE model in terms of F-score. The B-TransE model also addresses the shortcoming of the single model when it comes false negatives. At the cost of increasing the number of false positives, the B-TransE model manages to reach a recall of 0.67, an increase of 48 percentage points in absolute terms over the single TransE (Fake) model.

The results do not seem to correspond to those of Pan et al., particularly in the case of the single TransE (Fake) model. The reasons for this are primarily that a different dataset was used, and that the methodology implemented could not be obtained for replication, although attempts were made to follow the methodology presented as closely as possible. However, the improvements in performance for the B-TransE over the single TransE model are clearly significant. The differences in B-TransE performance may also be due, in part, to the difference in training set sizes between the two experiments, using 1000 articles compared to our 800 articles.

For our investigation, false positives are less costly than false negatives, since we would rather flag potentially fake articles for further investigation than spread unverified news. This means that the specificity of the classifier is less important than the sensitivity. The precision-recall curve is a useful tool to evaluate this trade-off. The precision-recall curve in Figure (4.1) shows a clearer picture of the precision in terms of recall for the B-TransE model. The classifier reaches around 10% in recall before it produces the first false positive. The AUPRC statistic for this curve is 0.77, which far exceeds the true positive proportion of 0.5 for a random classifier. There is also a near-plateau in terms of precision recall trade-off between recall values of 0.15 and 0.80, which allows the model to obtain higher recall without sacrificing too much in terms of precision, or in other words, without increasing the number of false positives by a significant amount.

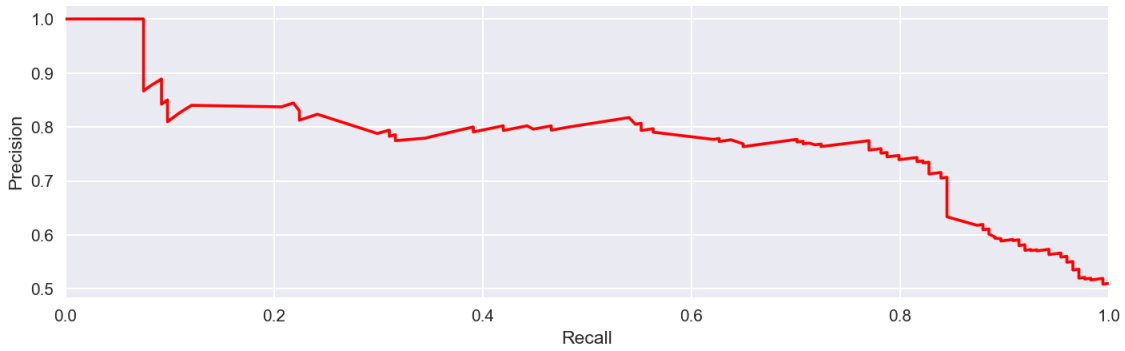


Figure 4.1: Precision-recall curve for the B-TransE model at thresholds between -0.12 and 0.08.

Since the TransE model trains embeddings only for entities and relations found in the training set, which is relatively small, it may be the case that randomly initialised embeddings are resulting in a large number of false negatives or positives for many triples in the training set. To investigate this, we removed all the triples from the test set which did not contain relations or entities previously found in the training set. This resulted in 70% of the test set being discarded. The results for the remaining 30% are shown in Table 4.2.

Model	Precision	Recall	F-Score
TransE (Fake)	0.83	0.49	0.62
B-TransE (English)	0.72	0.76	0.74

Table 4.2: 5-fold Cross-validated results from the evaluation of the remaining 30% of test set after filtering out unseen entities and relations

From Table 4.2, it is clear that both the single TransE model and the B-TransE model seem to be much more in line with the performance of the models presented by Pan et al. in terms of F-score, and in the case of the B-TransE model, the results come within 3 percentage points of the reported F-score by Pan et al. This shows that by avoiding random, uninformed predictions the model performs relatively well at discriminating the true news from fake news. In this setting, it is perhaps most prudent given the small training set to say that the other 70% of test articles are simply 'unknown'. It is unclear from the literature whether this was at all a consideration in the experiment by Pan et al.

4.2 Key Insights

4.2.1 2016 U.S. Presidential Election Data

The results from the 2016 U.S. Presidential Election data showed a large degree of variation in bias. We can see in Figure (4.2) that the B-TransE model correctly classified most of the extreme cases on both the true and fake partitions of the test set.

Figure 4.2: The top 30 (left) and bottom 30 (right) of the articles ranked according to the difference between the fake model bias and the true model bias.

When we look more closely at the content of the extreme cases, we are able to identify the maximum bias statements that were used for the classification decision.

Table (4.3) shows 5 of the top news articles identified as most likely to be fake, as well as the max bias triples that led to this classification decision.

Title	**Max bias triple**
Obama's gitmo board releases "high risk" explosive's expert, Al-Qaeda Trainer.	*(Barack Obama, release, risk)*
Obama made Christian Pastor pay for secret own ticket home after Iran got secret $1.7 billion ransom for secret release.	*(Christian, pay for, ticket home), (Iran, get, billion ransom)*
Careless Clinton Aide Kept 'Top Secret' State Department Info In Unsafe Locations.	*(Hillary Clinton, keep, state department info), (state department info, is in, unsafe locations)*
Which is extremely concerning seeing as how Obama has been known to recruit Muslim Foreign Service Officers through Jihad Conferences, as reported here at Judicial Watch.	*(Barack Obama, recruit, muslim)*
One of "Bernie's basement dwellers" on Fox showing off Bill Clinton rapist T-shirt	*(Bill Clinton, be, rapist)*

Table 4.3: Examples of articles classified as 'fake' by the B-TransE model which were in fact 'fake'. On the right, the triple with the maximum bias is higlighted

In many cases, it seems that single triples or near-duplicate variations of triples are the most common in scoring each article. The table shows that there is clearly some data loss, but also that some important information is retained, even in triples that are seemingly trivial.

In the first triple, we see that the model correctly extracts the most important information, but that the oversimplification of the tail entity leave the information somewhat ambiguous. The second and third triples shows how the overgeneration of Stanford OpenIE can work to the model's benefit, as we see that the model has extracted two equal max bias triples that tie two statements together. In particular, we see a clear link between the tail entity of the third article max bias triple (Hillary, keep, state department info), and then (state department info, is in, unsafe locations). The model has managed to successfully embed those statements together into a composite statement that can be verified. In the fourth article, the triple extracted seems to be appropriate and capture the main idea, although oversimplification in the tail entity could be refined once again.

The last triple clearly illustrates a large amount of information loss and the dangers of choosing a single triple to represent an article in aggregate. The triple does not mention Bernie Sanders, nor does it refer to signage on a T-shirt but simply puts out a statement that 'Bill Clinton is a rapist'. This is of course a statement which can be investigated, but it is not the main argument of the article, which refers to the act of Bernie Sanders supporters wearing said T-shirts. However, in most of the

cases highlighted (3 out of 5), we see that the max bias triple succeeds in capturing a succinct summary of the main argument in the article, and in almost all cases provides a triple which can be fact-checked or verified.

Table (4.4) looks at 5 of the articles identified as most likely to be true. In the first article, we can see a clear weakness of the Stanford OpenIE relation extraction framework. Here, the verb "lost" is not picked up as a relation, which causes it to move to the tail entity. When this entity is simplified, the verb is lost entirely. However, in dealing with simple single-word verb-mediated sentences, the model mostly succeeds in providing statements that capture the main ideas in each context, and also provides statements which are verifiable to a large extent.

The fourth and fifth articles illustrate trivial cases of a simple verb and word order that Stanford OpenIE can readily use to extract triples. In the second article, the max bias triple preserves the most important entities. However, the lack of named entities in the tail entity results in a large loss of information. Perhaps this would have been remedied if multiple rounds of extraction could be performed, first extracting (Icahn, say, Y), and then performing the extraction on Y to return (Donald Trump, better than, Clinton U.S. economy) to reduce the loss of information. Overall, it seems that the model has been able to provide the user with verifiable statements that capture the main idea in a particular news article, although the performance varies quite a lot depending on article construction and grammatical complexity.

Title	**Max bias triple**
Indiana Governor Mike Pence, the Republican nominee for U.S. vice president, lost another round in federal court on Monday	*(Pence, nominee for, U.S. Vice President)*
Icahn says Trump better for U.S. economy than Clinton.	*(Icahn, say, Donald Trump U.S. economy Hillary Clinton)*
President-elect Donald Trump will have an early Capitol Hill honeymoon with Republican majorities in both chambers of Congress when he takes office in January	*(Republican, is in, U.S. congress)*
Trump won the U.S. presidency with less support from black and hispanic voters than any president in at least 40 years	*(Donald Trump, win, U.S. presidency)*
Trump campaign says it raised $80 million in July: statement.	*(Donald Trump, raise, $ 80 million)*

Table 4.4: Examples of articles classified as 'true' by the B-TransE model which were in fact 'true'. On the right, the triple with the maximum bias is higlighted

Given that the model has been able to classify so many of the extreme cases correctly with triples that are often oversimplified, the question is then whether the model has been able to generalise an understanding of 'true' vs 'fake' using these simple

triples based on the TransE objective function. Alternatively, perhaps the model is simply creating bag-of-words counts of entity-relation bi-grams and building a frequency-based classifier. To address the former question, we first look at the top 4 relations used in the model to see how the entity and relation embeddings have been trained according to the objective function. The embeddings are visualised using PCA for dimensionality reduction, and the plot shows only the first two principal components. The goal is to see whether a valid decision boundary based on distance between fake triples and true triples has been created.

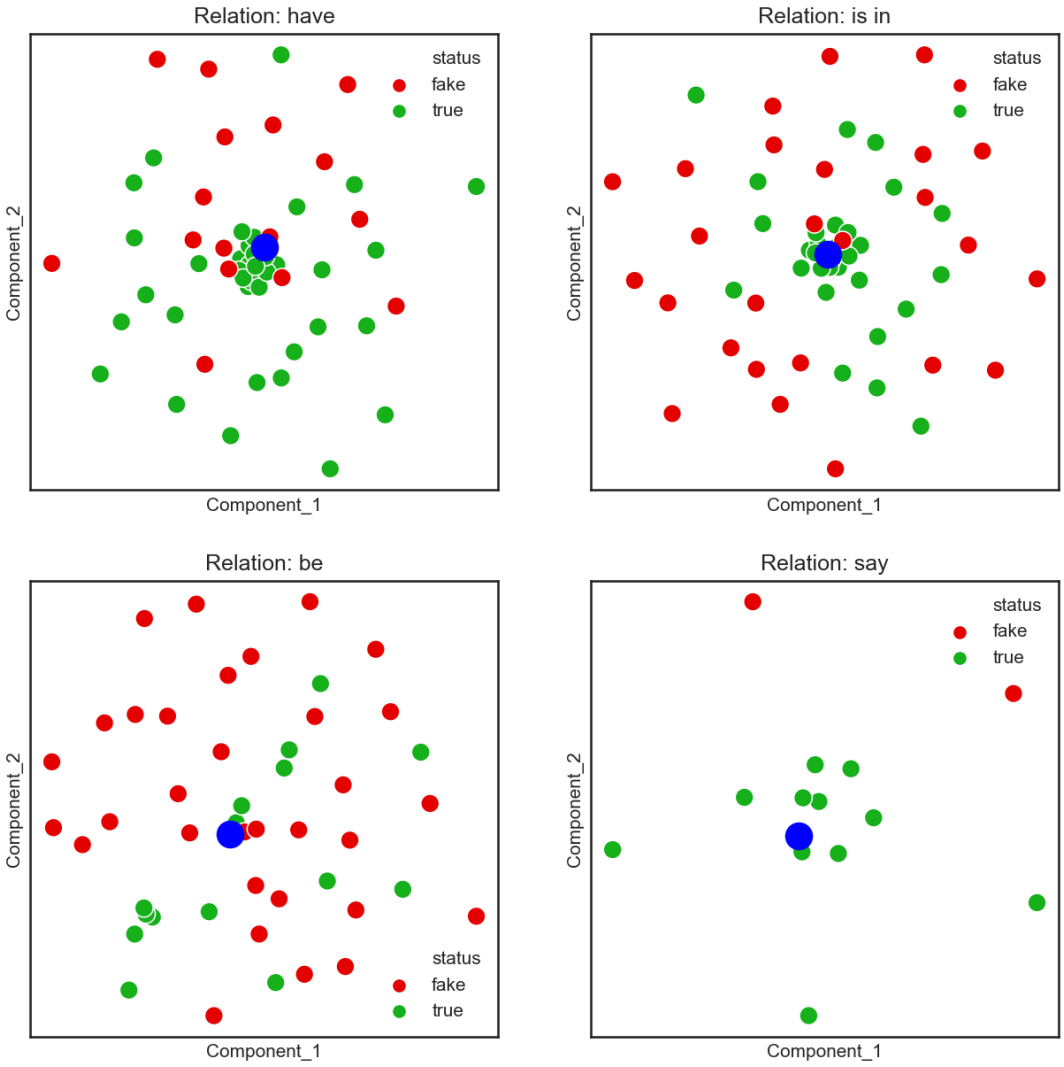


Figure 4.3: The top four relations are chosen from the test set (have, is in, be and say). Each dot denotes a triple and its position is determined by the difference $(h - t)$ between head and tail entity vectors. Since TransE aims to obtain $t - h \approx r$, the ideal situation is that there exists a single cluster whose centre is the relation vector, r. The statements from fake news are in red and the statements from the true news are represented by green dots. The larger blue dots indicate the ideal fit between the difference vector and the relation vector. The relation vector is obtained from the True TransE model embeddings [33].

The relations in Figure (4.3) show that the true dots tend to concentrate around

the true relation embedding and that the distance between the true relation embedding (blue dot) and the fake difference vector embeddings (in red) is generally greater than the distance between the true difference vector embeddings (in blue). The exception here seems to be the relation ”be”, which may indicate that this is a difficult relation to separate based on the triples. What this likely also indicates is that the simplification of all relations from ”is”, ”are”, ”become” to ”be” using lemmatization may have oversimplified these relations to a greater extent than necessary - leading to a loss of information. The other important consideration is the aggregation from the triple to the article level. It may be the case that individual triples within fake articles are closer to triples embedded from true articles or vice-versa e.g. *(Hillary Clinton, be, Planned Parenting Supporter)* is from a fake article, but is a true statement of Hillary's allegiance.

Given the top four relations, we can conclude that the model has captured mostly simple statements. This is in some sense a useful common sense check on the B-TransE model, for e.g. in the relation 'is in', the model confirms the following triples as being close to the truth: *(China, is in, Asia), (Russia, have, Crimea) and (Donald Trump, have, Hollywood Walk of Fame).* However, in fake triples, the simplification of entities produces statements which are more difficult to verify by eye, including *(US failures, is in, Libya), (Huma, be, Judicial Watch) and (Jeff Sessions, have, Republicans).* A more complete view of these is provided in Appendix 1.

This is only a partial answer, since if we look closer at the most common bi-grams in the training set for labelled 'True' and 'Fake' articles shown in Table (4.5), we notice a few interesting patterns. Firstly, it seems that there is a substantial amount of overlap, with 3 out of the top 10 entity relation pairs shared between the two classifications, which suggests that the TransE model has been effective in determining a distance-based decision boundary. At the same time, it should be noted that 'Hillary Clinton' appears 4 times in the top 10 entity-relation pairs in the 'fake' news category, whilst only appearing twice in the 'true' news. This points to a possible frequency based classification where any triple containing 'Hillary Clinton' as the head entity with any of the most frequent relations (e.g. be, have) will be strongly biased towards 'fake' classification in the model. This also perhaps indicates a strong bias in 'fake' articles towards describing Hillary Clinton and her actions, whilst the 'true' articles revolve around a more balanced group of individuals in the election discourse.

Entity	Relation
Hillary Clinton	hold
Donald Trump	have
Florida	register
Paul	be
Donald Trump	be
Paul Ryan	be
Washington	hand over
Hillary Clinton	have
Gary Johnson	frame
Wisconsin senate	could gain

(a) Bi-grams from True articles

Entity	Relation
Tony	be
James Clapper	was testifying
Obama	was in
Donald Trump	be
Hillary Clinton	be
Hillary Clinton	select
Hillary Clinton	have
Juanita	be
Hillary Clinton	was at time
Donald Trump	have

(b) Bi-grams from Fake articles

Table 4.5: The top ten (entity, relation) bi-grams from the 'True' articles in (a) and 'Fake' articles in (b) from the training set

4.3 Fake News Generation

An interesting question to pose for knowledge embedded models is whether they can act as alternative fact generators. To understand what the model has come to think of as 'fake' in the context of the 2016 U.S. Presidential Elections, we look at the top 5 nearest neighbours for a combination of entities and relations that frequently appear in fake news articles. These are (Donald Trump, be), (Donald Trump, have), (Hillary Clinton, be) and (Hillary Clinton, have).

Alternative Facts

Below are examples of alternative facts generated by the model. It seems that in some cases, the embeddings do mirror alternative facts from fake news articles, especially in the case of Hillary Clinton with negative suggestions such as (Hillary Clinton, have, lie) or (Hillary Clinton, have, step down) or (Hillary, be, arms dealer libyan weapons). The generated examples involving Donald Trump seem to be also be much more neutral in nature e.g. (Donald Trump, be, U.S. Senate) and (Donald Trump, have, reform). This seems to suggest that these fake news articles habitually build up Donald Trump through a description of his actions and responsibilities and paint Hillary Clinton in a particularly negative light.

Candidate: ['donald trump', 'be']
Nearest Neighbours: [['karl ove knausgrd', 'president vladimir putin', 'us senate', 'scranton', 'denis', 'ecuador', 'abc', 'treasury', 'howard stern']]

Candidate: ['donald trump', 'have']
Nearest Neighbours: [['reform'], ['earnest'], ['islamic san'], ['crooked harry'], ['carolina governor haley'], ['zuckerberg'], ['jack'], ['rocky roque de la'], ['willie'], ['republican']]

Candidate: ['hillary clinton', 'be']
Nearest Neighbours: [['current', 'xi jinping', 'arms dealer libyan weapons', 'ecuador', 'us senate', 'insane', 'fredrik', 'zika','democratic strategist', 'putin']
]

Candidate: ['hillary clinton', 'have']
Nearest Neighbours: [['carolina governor haley'], ['step down'], ['mike republican mike rogers'], ['islamic san'], ['willie'], ['isis'], ['muslim american iraq'], ['lie'], ['paul'], ['facebook us']]

"True" Facts

The model generates some interesting facts surrounding both Hillary Clinton and Donald Trump. The true facts are more inclined to paint Donald Trump with a negative brush (violence, communist sympathizer), although a great deal of noise is also present in the tail entities. For Hillary Clinton, we see a shift in sentiment towards positive entities such as "morally choice" and "support". For both of them we see the "possiblepotusy" tail entity which refers to possible President of the U.S — a statement that is indeed verified.

Candidate: ['donald trump', 'be']
Nearest Neighbours: [['violence'], ['christoph blocher'], ['philippines obama'], ['democratic national convention philadelphia'], ['communist sympathizer'], ['us house wisconsin washington'], ['menlo park'], ['possibilpotusy'], ['thiel'], ['assad']]

Candidate: ['donald trump', 'have']
Nearest Neighbours: [['osce united states'], ['united mine workers america'], ['schumer'], ['tony'], ['factbox'], ['upflynnaval'], ['tpresident tayyip erdogan'], ['naked'], ['scranton'], ['robby mook']]

Candidate: ['hillary clinton', 'be']
Nearest Neighbours: [['christoph blocher'], ['possibilpotusy'], ['marco gutierrez'], ['assad'], ['pew'], ['isis mosul iraq'], ['planned parenthood'], ['morally choice'], ['support'], ['faux fight']]

Candidate: ['hillary clinton', 'have']
Nearest Neighbours: [['upflynnaval'], ['washington party'], ['julian assange'], ['scranton'], ['schumer'], ['tony'], ['osce united states'], ['japan pm abe'], ['core'], ['opposition']]

Overall, the above has shown that there is a great deal of noise in the generation of alternative facts, which indicates that the model may not perform well on generating plausible fake news from only 1600 articles.

4.3.1 Swedish News Data

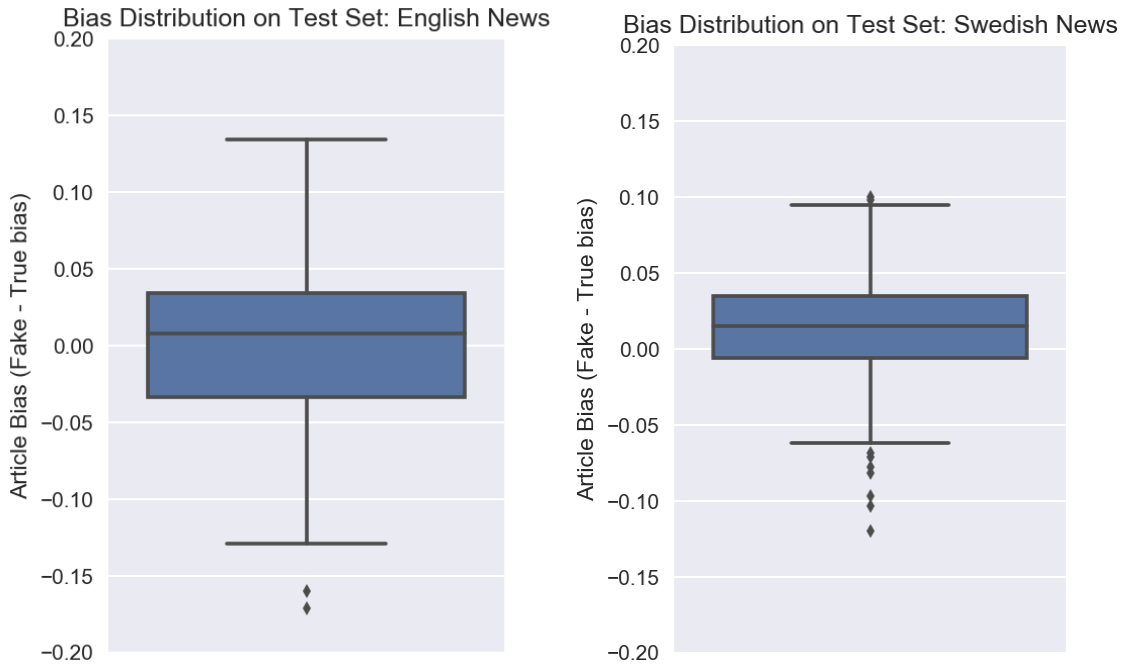
4.3.1.1 Classification

From the 1441 articles in the Swedish test set, the B-TransE model classifies 798 articles as 'true' and 643 articles as 'fake'. The proportion of fake news predicted seems out of line, given our expectation that around 1 in 3 articles may be fake [14]. It should however be noted that there was a large degree of homogeneity in the content of Swedish news, meaning that if a particular event was reported by a Swedish newspaper, it could be replicated nearly word-by-word by another Swedish newspaper, since they are extremely likely to share the same primary source for event information from the United States. This means that incorrect predictions by the model are amplified somewhat as for e.g. 50 wrong predictions could in fact reduce to a single article being wrongly predicted.

4.3.1.2 Bias Distribution

The boxplots in Figure 4.4 show that the model trained on English data sees the U.S. News test dataset as having much larger variation in terms of bias differentials than the Swedish news dataset. The median article seems to lie around the uncertain outcome of 0 for the English data, and around 0.02 for the Swedish data. The max bias differential in the Swedish dataset is 0.09 compared to 0.13 in the English data. The difference in the minimum bias differential is even more pronounced at -0.12 for the Swedish data and -0.17 for the English data.

This shows that the model considers the Swedish data to be more balanced and less extreme when compared to the English news dataset. There could be many plausible explanations for this, such as the fact that Swedish coverage of a foreign election could be much more balanced than in the United States where there are political points to be scored. The narrow inter-quartile range in the Swedish dataset also points to increased uncertainty, with a large proportion of articles close to a zero bias differential. It may not seem like the differences are particularly large, but in models for fake news detection the extremes and outliers are useful for the prediction of fake news, as these articles typically represent a tiny minority of all the news coverage. Overall, if we can understand the differences in distribution between extreme values in both languages, this could potentially help us to better understand the outliers in Swedish data.



(a) Article Bias Boxplot: English data **(b)** Article Bias Boxplot: Swedish data

Figure 4.4: Boxplots showing distribution the different in fake bias - true bias at an article level for both the English data in (a) and the Swedish data in (b)

4.3.1.3 Extreme Cases

Table (4.6) gives an indication of the articles for which the B-TransE model had the highest confidence in prediction for 'fake'. Though the statements classified as 'fake' are difficult to verify thoroughly, each of these triples is cross-referenced to the closest triple found in the training set to decide whether the classification was correct according to the knowledge graph. One interesting observation from these articles is that there seems to be an above average representation of Donald Trump in the articles chosen, which differs from the English model where 'Hillary Clinton' was the most prominent candidate for 'fake' articles.

The 1st, 2nd and 5th articles were indeed verified as 'fake' by cross-referencing these articles to the 'Fake' English language articles in the training set. However, the 3rd and 4th articles were found in both the 'True' and 'Fake' articles of the training set, and thus these classifications are incorrect. The 2nd article illustrates a situation where the translation from Swedish has created confusion, as the word "buades" was wrongly translated to "evicted by" instead of "booed out", distorting the meaning. Overall, the model has been successful at identifying most of the fake articles, and in cases where there is uncertainty (appearing in both 'True' and 'False' knowledge graphs), the model flags these articles, which is a useful signal to fact-checkers to dig deeper.

Title	**Max bias triple**
Robert De Niro wants to beat Donald Trump.	*(Robert Dinero, beat, Donald Trump)*
Trump was evicted by Catholics	*(Donald Trump, was evicted by, Catholics)*
Bush Leaves Show "Today" after the stir on the recordings with Donald Trump.	*(Bush, leave, show)*
Jessica Drake is the 11th woman to blame Trump for assault	*(Jessica, blame, Donald Trump)*
Michelle Obama attacked Trump. She has said that she is looking forward to a quieter life, after the time as president wife.	*(Michelle Obama, attack, Donald Trump)*

Table 4.6: Examples of articles classified as 'fake' by the B-TransE model for the Swedish news data. On the right, the triple with the maximum bias is higlighted

In Table 4.7, we once again attempt to verify these statements by cross-referencing with the training set. The first 3 articles could be verified in the 'True' news articles, whilst the last two did not appear in either the 'True' or 'Fake' training sets. In the final max bias triple the phrase "backa om" is translated to "backs over", when it should be "moves away from" or "reverses". This could be another reason why this article could not be found in the training data. On the whole, the model does well to determine the most informative triple from each article, and the model has managed to verify these triples, when the translation is of adequate quality.

Title	Max bias triple
Michael Moore releases Trump movie.	*(Michael Moore, release, Donald Trump)*
Trump is in Washington to discuss foreign policy.	*(Donald Trump, is in, Washington)*
Buffett releases information from his tax return and urges Trump to do the same.	*(Buffett, urge, Donald Trump)*
Asia's Trump in Flirt with China.	*(Asia, have, Donald Trump), (Donald Trump, is in, flirt China)*
Trump backs over Putin. The US Republican presidential candidate Donald Trump earlier celebrations of Putin at a campaign meeting in Nevada. - I do not love.	*(Donald Trump, backs over, Putin)*

Table 4.7: Examples of articles classified as 'true' by the B-TransE model for the Swedish news data. On the right, the triple with the maximum bias is higlighted

4.4 Biases

It is important to acknowledge that there are inherent biases in the methodology presented in this research project. Some of the key biases to be recognised when interpreting the results are:

1. The labeled training set from Politifact was scored based on experts reviewing articles, but the categories were not simply "true" or "false", but rather more nuanced including "mostly false", "mostly true" and "pants on fire". The source of the ISOT dataset does not disclose which categories were chosen when constructing the fake news dataset.

2. When the TransE model is trained, the model may overfit to individual statements rather than being able to assess the article as a whole because individual statements inside a fake news article need not only be false, and vice-versa for a true news article. Thus, the choice of aggregation function is in itself a form of bias that influences the effectiveness of the classifier. This is a clear distinction between applications to raw news articles versus a knowledge base where individual triples have been verified.

4.5 Model limitations

It is important to highlight some of the known methodological limitations imposed by the models used, including:

1. The TransE model is a global embedding model, i.e. the embedding of entities and relations are context-independent, i.e. the entity embedding of ”Trump” is identical when talking about the TV-show ”The Apprentice” and the U.S. presidential elections, even though these contexts clearly have different likelihoods of containing certain triples.

2. The TransE model relies on embedding entities and relations that are found in the training set. In practice, the model does not handle unseen entities and relations, and simply assumes that the embedding is random (i.e. uninformative).

3. The TransE model is most appropriate for modelling 1-1 relations, and not many-to-1, many-to-many or 1-to-many relations. This is a drawback as more complex relationships cannot be explored using this model.

4. The Stanford OpenIE package fails in many cases to adequately deal with long sequences containing multiple verbs. One of the main reasons for this is that many entities are not recognized as named entities (e.g. persons, locations and organizations).

5. The Google NMT model, whilst constantly improving, does not do a faultless job of translating English to Swedish text, especially when it comes to local idioms and expressions or translating short sequences. As we have seen, this may lead to a distorted meaning that acts as hurdle to the generalisation of the model.

6. The knowledge graph approach focuses on the content and thus has a literal understanding of statements. Therefore, other linguistic aspects such as hypothetical statements, irony or the use of negation are not readily detected. However, as this was not the primary focus, it is seen as outside the scope of this project to deal with this implication.

5

Conclusion

5.1 Summary of goals and contributions

In this research project, the task of content-based fake news detection was addressed. The first stated goal centered around scoring statements from news articles using a knowledge graph embedding model to replicate the results of a novel paper by Pan et al. As far as the author knows, the aforementioned paper contains the first attempt to detect fake news using knowledge graphs and raw news articles. This goal was achieved to some extent with an F-score obtained within 3% of the original paper for the B-TransE model given a set of important assumptions. However, due to a lack of information on the pre-processing pipeline and no access to the dataset, the results of both the single Fake TransE model and the B-TransE model could not be reproduced consistently.

The B-TransE model also managed to achieve an F-score of 0.74 using only a relatively small dataset of raw news articles (1600), which translates into useful applications for languages with small labeled training sets. This meant that the knowledge graph approach was quite successful at detecting fake news in incomplete knowledge graphs populated mostly by simple facts. Thus, a large existing knowledge base is not required to obtain a similar level of performance as smaller contextually appropriate temporal model. By proceeding in this way, the model's data requirements, complexity and training times can all be kept relatively low.

The project also managed to deepen the understanding of the B-TransE model predictions by explaining the results, in line with a trend in the field to move towards explainable AI systems and away from black box methods [35]. This was also an explicit objective mentioned in the 'Future work' section of the Pan et al. paper. An unexpected realisation from this particular project was that although the B-TransE model had only a pure fact-based construction and focus, it was inevitably influenced by the stylistic features from the news articles when making predictions. This brings up the question of whether these two approaches are truly mutually exclusive, or whether hybrid models are an inevitable consequence of the nature of the development of fake news, both in style and content.

The final goal was to construct a reference dataset for the Swedish language using

a transfer learning approach. This was not really achieved to a large extent as the model showed that the distribution of bias differences was quite different between English and Swedish, suggesting that further investigation may be needed to increase the confidence in labelling. Having said that, the model was able to correctly classify the majority of the top 5 'Fake' and 'True' predicted news articles. The model also illuminated some of the key biases that have been ported from the English language model. Hopefully, this could be explored further and used as a starting point to understanding what defining fake news looks like the Swedish context.

Overall, it became clear through detailed analysis that the fake news problem is much more nuanced and complex than first expected. To make use of knowledge embeddings in a reasonable way, the amount of pre-processing required is tremendous, and the trade-off between extracting useful information for fact-checking and losing valuable descriptive data is clear. The project also showed that even a knowledge embedding model with a fairly transparent premise can be difficult to interpret when it relates to entities and relations that are often unclear and complex.

5.2 Ethical considerations

One important ethical question that keeps coming up in this problem space is whether news should be screened and filtered, and if so, at which point it should be done. Is it up to the incentivisers and aggregators like Facebook and Google, or the journalists from the news outlets, or even the readers themselves? It is the author's point of view that rather than opt to withhold information from the public, it should be the focus to empower them with the likely veracity of a particular piece of information.

At the same time, it should be the goal to equip journalists and other trained fact-checkers with a suite of tools (such as the model presented in this project) that decrease the likelihood of propagating false information. It is important to keep in mind that these models should be used as a complement to increase the efficiency of current fact-checking procedures through explainable automation for reliable news sources. It is therefore imperative that a human be kept in the loop to make the final decision on whether the model outcome is reasonable or not - as this also implies an accountable party when mistakes are made.

Additionally, third parties such as Google and Facebook that unwittingly create incentive structures around publishing what is essentially 'clickbait' content should acknowledge their role in spreading unverified news stories, whilst also aiming to caution and educate users rather than restrict access to fake news information. The Silicon Valley startup, FactMata, has set out to do just that, by providing available facts on any statement expressed in news media or social media so that the reader can make an informed judgment. Essentially, end-users should become the final fact-checkers when all other defences are thwarted [9].

5.3 Future developments

Apart from the main goals, the knowledge graph embedding model could also have auxiliary intended uses, including:

1. Given a source, return a reliability score by evaluating all given statements from that source and verifying as many as possible. The proportion of verified facts is one such metric.

2. Given a source and a statement, use the reliability score of the source to incorporate new evidence and update the knowledge graph embeddings to reflect newly verified information.

In future, this research could be extended in a number of ways.

1. Firstly, it is clear that information loss due to a small training set size is a very relevant problem for this approach. To remedy this, one could look at using models that deal with unseen entities and relations, such as the model proposed by Xiong et al. [34].

2. A more advanced knowledge embedding model, DOLORES, has been proposed as shown in Figure 5.1. It uses ”deep contextualised embeddings” to infer missing links in the knowledge graph using a sequential neural network that learns both contextual and non-contextual information. It does this by embedding the entities together with their relations rather than the triples themselves. There are two main components of the DOLORES framework. The first is a path generator which creates sequences of entity-relation pairs that represent a sample of neighbourhood exploration in the knowledge graph. The second is a deep recurrent neural network, which is trained on the entity relation pair sequences. The embeddings trained for each entity and relation represent non-contextual information, whilst the weights of the sequential model help to capture temporal information in the order in which it is presented to the model. Thus, this would allow the model to correct a previously ”fake” story that has been corrected by a particular news source. The experimental results in triple classification on two reference knowledge bases, WN11 and FB13 show that DOLORES embeddings improve the classification performance by 1.4% to 88.4% over the state-of-the-art [28].

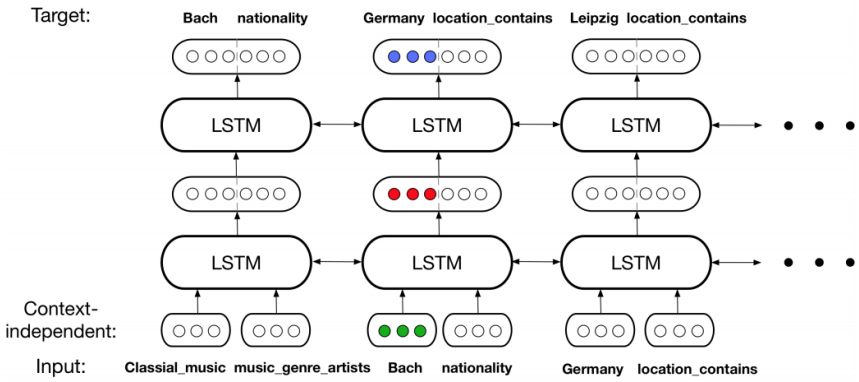


Figure 5.1: Unrolled RNN architecture of the DOLORES model. First, the entity-relation chains are generated from random walks in the local neighbourhoods of the knowledge graph, and then used as input to the bi-directional LSTM. These entity and relation vectors are concatenated after being initialised and are then fed to the LSTM. The model learns the contextualised representation of each entity from the weights of deeper network layers (in red and blue). This figure was taken from [28].

3. A model capable of handling streaming data could be constructed using the same lightweight principles of limited time and event domain. This would help to reduce the need to retrain the model each time retrospectively for an event domain in the past.

4. Models that look at contested topics are also an interesting avenue to pursue. By following the two strongest anti-poles in a controversial topic setting, this could be used rather than extensive background knowledge to verify arguments on both sides of the aisle. This is especially crucial in contexts where no labeled dataset exists.

Bibliography

[1] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9. doi: 10.1002/spy2.9. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9`.

[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In Issa Traore, Isaac Woungang, and Ahmed Awad, editors, *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham, 2017. Springer International Publishing. ISBN 978-3-319-69155-8.

[3] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Working Paper 23089, National Bureau of Economic Research, January 2017. URL `http://www.nber.org/papers/w23089`.

[4] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1034. URL `https://www.aclweb.org/anthology/P15-1034`.

[5] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. *CoRR*, abs/1810.01765, 2018. URL `http://arxiv.org/abs/1810.01765`.

[6] Akim Bassa, Mark Kroll, and Roman Kern. GerIE - an open information extraction system for the German language. *Journal of Universal Computer Science*, 24(1):2–24, Jan 2018.

[7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/`

5071-translating-embeddings-for-modeling-multi-relational-data.pdf.

[8] C. Buntain and J. Golbeck. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215, Nov 2017. doi: 10.1109/SmartCloud.2017.40.

[9] Robert Dale. NLP in a post-truth world. *Natural Language Engineering*, 23 (2):319–324, 2017. doi: 10.1017/S1351324917000018.

[10] Florence Davey-Attlee and Isa Soares. The fake news machine: Inside a town gearing up for 2020, 2019. URL https://money.cnn.com/interactive/media/the-macedonia-story/.

[11] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610, 2014. URL http://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf. Evgeniy Gabrilovich Wilko Horn Ni Lao Kevin Murphy Thomas Strohmann Shaohua Sun Wei Zhang Geremy Heitz.

[12] Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. OpenKE: An open toolkit for knowledge embedding. In *Proceedings of EMNLP*, pages 139–144, 2018.

[13] Karen Hao. Ai is still terrible at spotting fake news, Oct 2018. URL https://www.technologyreview.com/s/612236/even-the-best-ai-for-spotting-fake-news-is-still-terrible/.

[14] F. Hedman, F. Sivnert, and L.M. Neudert. News and Political Information Consumption in Sweden: Mapping the 2018 Swedish General Election on Twitter. *COMPROP*, 2018.

[15] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558, 2016. URL http://arxiv.org/abs/1611.04558.

[16] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep Video Portraits. *ACM Transactions on Graphics 2018 (TOG)*, 2018.

[17] Andreas Klintberg. Train an swedish dependency parser for stanford corenlp. https://github.com/klintan/corenlp-swedish-depparse-model, 2017.

[18] Denis Krompaß, Stephan Baier, and Volker Tresp. Type-constrained representation learning in knowledge graphs. *CoRR*, abs/1508.02593, 2015. URL http://arxiv.org/abs/1508.02593.

[19] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68(8): 855 – 859, 2015. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2015.02.010. URL http://www.sciencedirect.com/science/article/pii/S0895435615001067.

[20] Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. Content based fake news detection using knowledge graphs. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*, pages 669–683, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00671-6.

[21] Kasey Panetta. Gartner top strategic predictions for 2018 and beyond, 2018. URL https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[23] 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*, abs/1707.03264, 2017.

[24] Victoria L. Rubin. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010. doi: 10.1002/meet.14504701124. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/meet.14504701124.

[25] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.*, 45(4):427–437, July 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.03.002. URL http://dx.doi.org/10.1016/j.ipm.2009.03.002.

[26] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. ISSN 0036-8075. doi: 10.1126/science.aap9559. URL https://science.sciencemag.org/content/359/6380/1146.

[27] Kathleen Walch. Knowledge graph applications in the enterprise gain steam, Dec 2018. URL https://searchenterpriseai.techtarget.com/feature/Knowledge-graph-applications-in-the-enterprise-gain-steam.

[28] Haoyu Wang, Vivek Kulkarni, and William Yang Wang. DOLORES: deep contextualized knowledge graph embeddings. *CoRR*, abs/1811.00147, 2018. URL http://arxiv.org/abs/1811.00147.

[29] Xuepeng Wang, Kang Liu, and Jun Zhao. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 366–376, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1034. URL `https://www.aclweb.org/anthology/P17-1034`.

[30] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *the 24th ACM SIGKDD International Conference*, pages 849–857, 07 2018. doi: 10.1145/3219819.3219903.

[31] Webhose. Swedish news articles - a free public dataset, 2016. URL `https://webhose.io/free-datasets/swedish-news-articles/`.

[32] Thomas Wolf. Neuralcoref 4.0: Coreference resolution in spacy with neural networks. `https://github.com/huggingface/neuralcoref`, 2017.

[33] Han Xiao, Minlie Huang, Yu Hao, and Xiaoyan Zhu. Transg : A generative mixture model for knowledge graph embedding. *CoRR*, abs/1509.05488, 2015. URL `http://arxiv.org/abs/1509.05488`.

[34] Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. One-shot relational learning for knowledge graphs. *CoRR*, abs/1808.09040, 2018. URL `http://arxiv.org/abs/1808.09040`.

[35] Carlos Zednik. Solving the black box problem: A general-purpose recipe for explainable artificial intelligence. *CoRR*, abs/1903.04361, 2019. URL `http://arxiv.org/abs/1903.04361`.

[36] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *ArXiv*, abs/1905.12616, 2019.

[37] Xinyi Zhou and Reza Zafarani. Fake news: A survey of research, detection methods, and opportunities. *CoRR*, abs/1812.00315, 2018. URL `http://arxiv.org/abs/1812.00315`.

[38] Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. Fake news detection via NLP is vulnerable to adversarial attacks. *CoRR*, abs/1901.09657, 2019. URL `http://arxiv.org/abs/1901.09657`.

A

Appendix 1

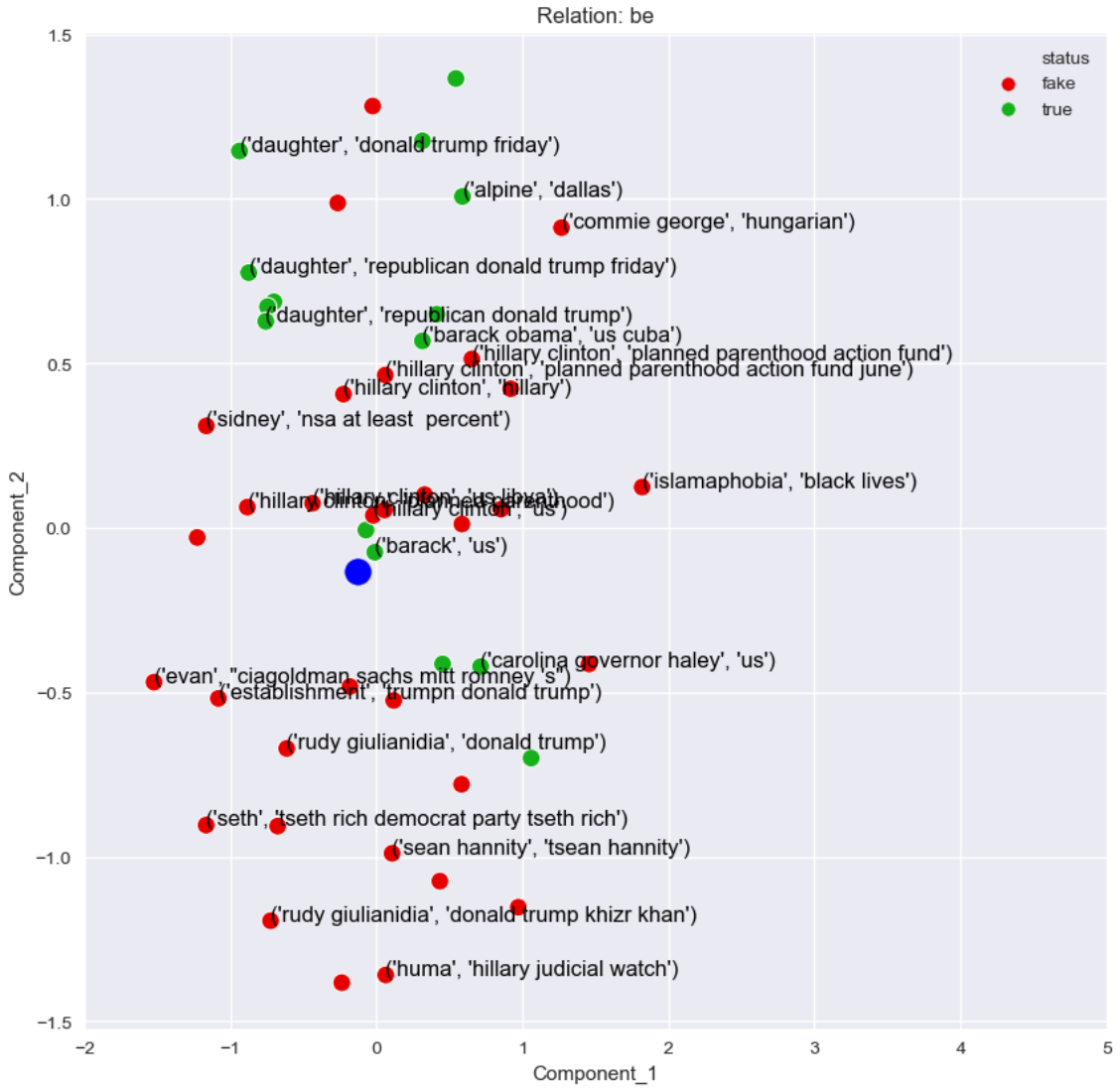


Figure A.1: Full representation of Figure 4.3 for the relation 'be'

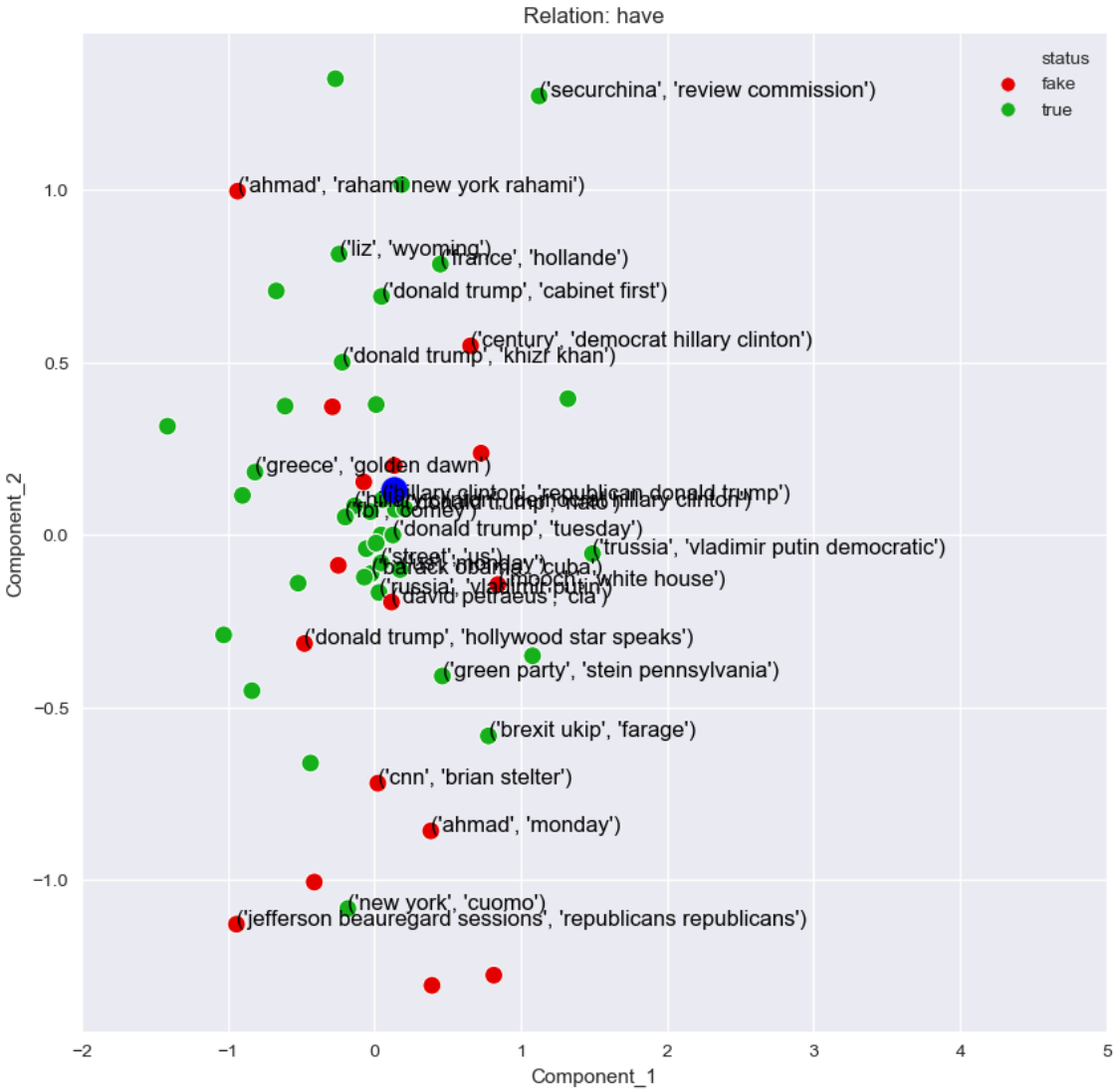


Figure A.2: Full representation of Figure 4.3 for the relation 'have'

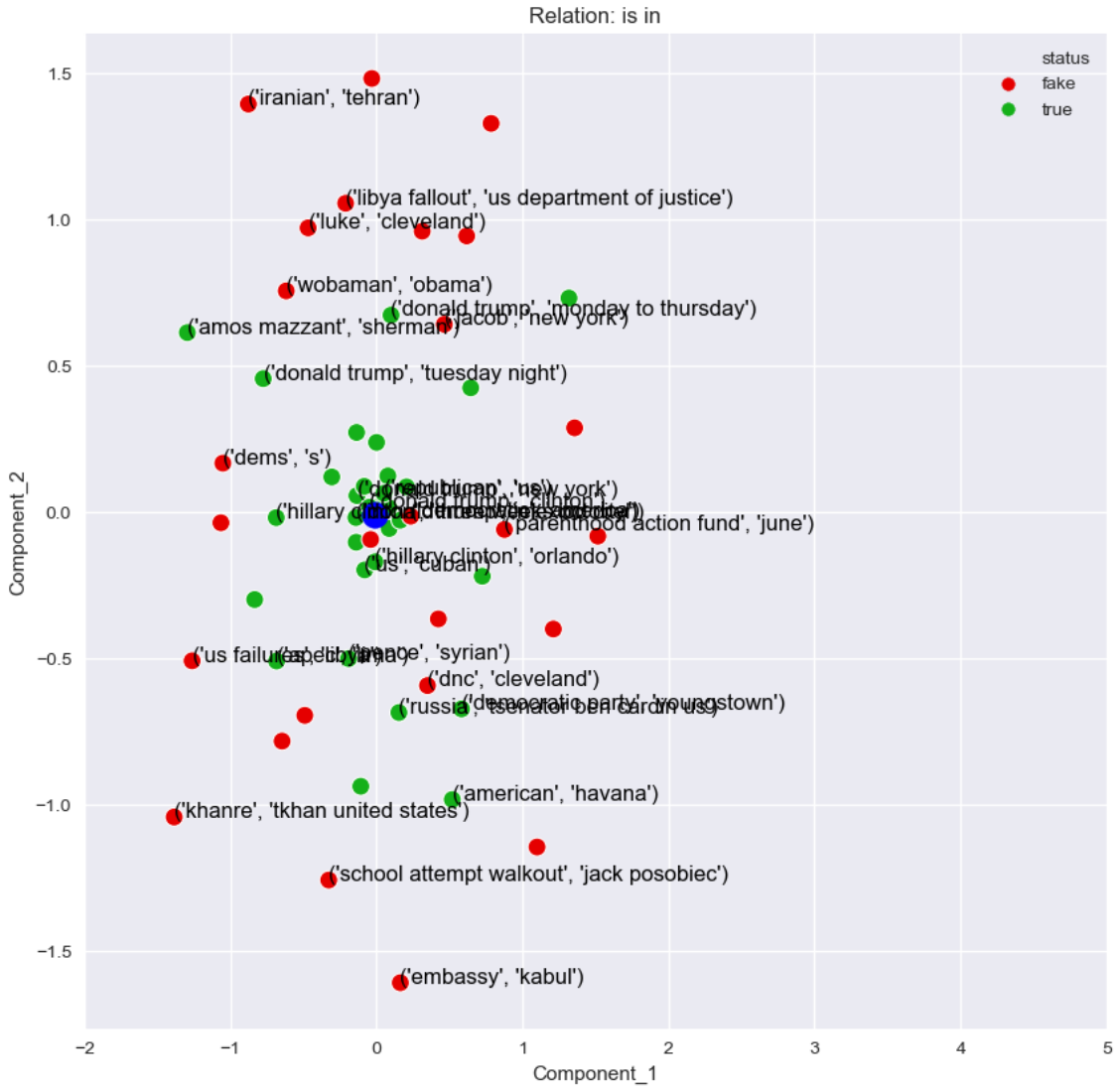


Figure A.3: Full representation of Figure 4.3 for the relation 'is in'

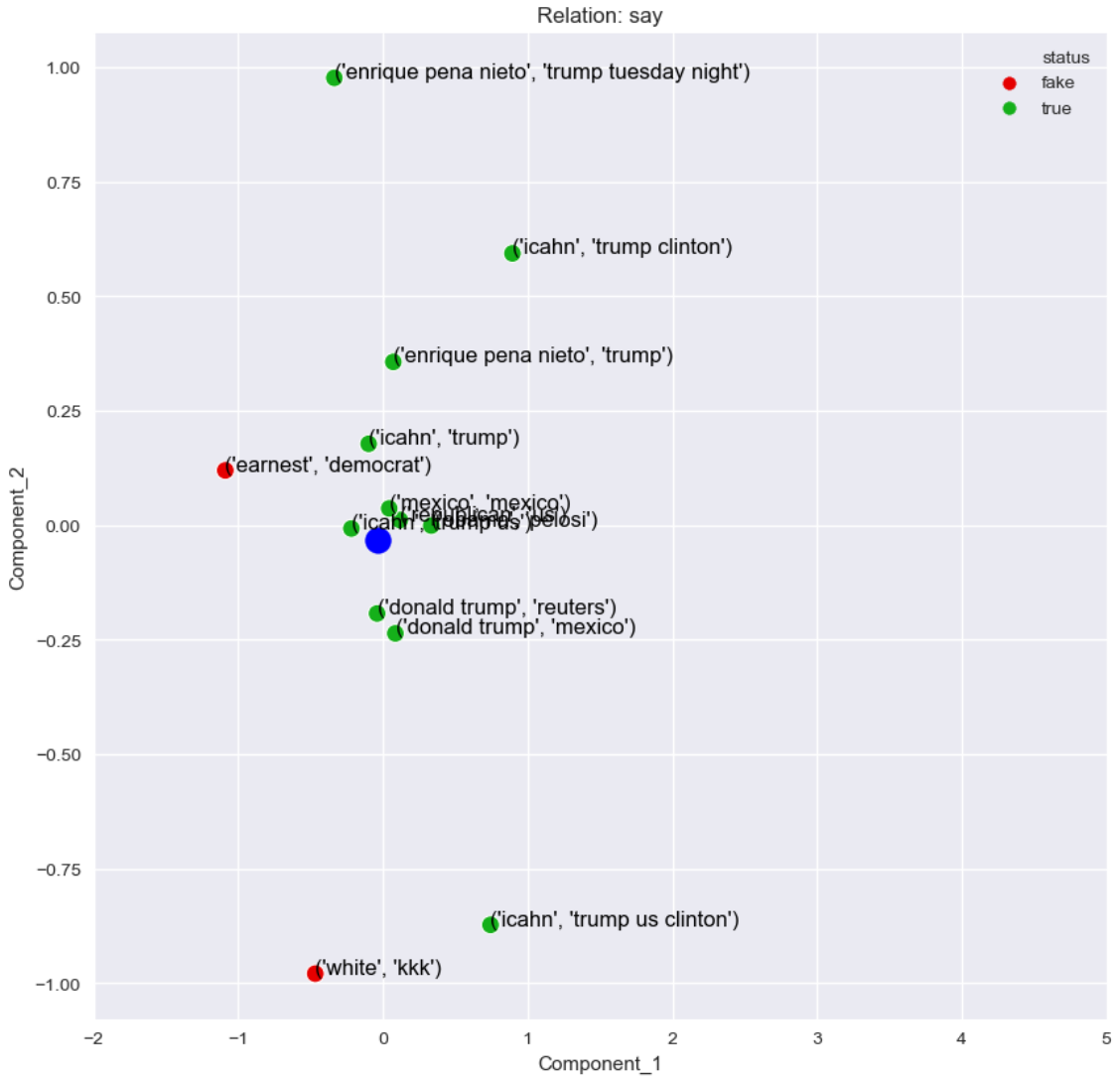


Figure A.4: Full representation of Figure 4.3 for the relation 'say'