Through the Limits of Newspeak: an Analysis of the Vector Representation of Words in George Orwell’s 1984

I. Dunđer* and M. Pavlovski*

*Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, 10000 Zagreb, Croatia
ivandunder@gmail.com, mpavlovs@ffzg.hr

Abstract - The era of fake news, media manipulation and information wars has been beneficent to the lasting fame and continuous acclaim of George Orwell’s 1984. The novel, published in 1949, influences to the present day the terminology of various political and social analysts through the use of its fictional language called “Newspeak”. The question arises – can the inner connections of the concepts present in Orwell’s 1984, when analysing the text on a semantic level of the words in their contextual environment, be used to further the understanding of the inner-workings of the novel’s language itself? More specifically, the aim of this paper is to examine whether a reader without knowledge of the subject of a fictional work of art, exemplified by Orwell’s 1984, could gain deeper comprehension of the text just from analysing word vector representations, without the use of external resources and only through an overview of the established similarities on the semantic level of words in a given text. In fact, word vector representations, as a form of word embeddings in a vector space model, are a machine learning technique sometimes applied in natural language processing, which attempts to identify semantically similar words.

Keywords - natural language processing (NLP); word vector representation; word2vec; word embeddings; vector space model; George Orwell’s 1984; dystopia; Newspeak; machine learning; information and communication sciences

I. INTRODUCTION AND MOTIVATION

George Orwell’s novel 1984 has had a wide influence on the terminology of various political and social analysts to the present day. Its fictional language “Newspeak”, as well as the Orwellian definition of concepts such as war or power are all relatable to the present era of fake news, media manipulation and information wars.

When analysing the text of Orwell’s 1984 on a semantic level of the words in their contextual environment, can the inner connections of its concepts be used to further the comprehension of the inner-workings of the language itself?

Could a reader, without any knowledge of the subject of a novel, play or other fictional works of art, gain a deeper understanding of the text just from a survey of words that were identified as semantically most similar by vector representations of words?

These vector representations on the semantic level, more specifically, an overview of the established similarities on the semantic level of words in a given text, could aid students to investigate the text without any topic-related knowledge or the use of additional resources. The application of such a method on a complex, literary text, could be a proof of concept for all other possible applications on simpler texts.

This approach, when applied with natural language processing techniques, could, therefore, enhance methods of text examination in literary studies, higher education and the academic curricula itself.

II. MOTIVATION

The aim of this paper is to study, on the example of Orwell’s 1984, if a reader could comprehend the meaning of the text just from a brief survey of its “semantic network”, the vector representation of words, without the use of any other resources, and only through an overview of the established similarities between words in the text.

Word vector representations, as a form of word embeddings in a vector space model, are a machine learning technique sometimes applied in natural language processing, which attempts to discover semantically comparable words.

This paper is a proof of concept – if one can successfully examine the “semantic network” of a fictional work of art, and in this case a complex novel with various important concepts and a whole fictional language used by the characters and invented specifically for it, then one could presumably apply this approach on less-complicated, non-fictional texts for various analytical purposes.

III. RELATED WORK

The terminology of Orwell’s 1984 has attracted a lot of attention and was examined by a wide range of scientists in higher education, not only from the field of literary studies but also from politics, media, information and communication sciences, linguistics and sociology, just to name a few.

In recent years, publications concerning the use of Orwellian terminology in politics and in the media has
been on the rise due to the so-called era of fake news. In relation to language, between ideology and power in Orwell’s 1984, [1] asserts that “ideology gains shape and manifests itself only through a semiotic system that is called “language”, and thus, whoever controls the language, controls the ideology, and consequently the power structures of a given society. By devising the concept of Newspeak in 1984, Orwell manifests his profound awareness of the relationship between language and the power relations in society”. Reference [2] uses the example of the novel’s famous quote “The object of power is power”, to reveal “tautological symbolic lack [...] as an unerring point of convergence between the totalitarian state envisioned by Orwell and the modern social democratic structures” – a claim that the paper exemplifies by examining British Prime Minister Theresa May’s tautology “Brexit means Brexit”. “Inspired by Orwell’s chilling account of brainwashing, propaganda and the obliteration of the lines between fiction and truth...”, [3] discusses the topic of fake news, alternative facts and “[...] the challenges that the advent of fake news organisations and the dissemination of alternative facts as truth pose to deliberative civics education, specifically how the trend of a diminishing space of truth and fact undermines efforts to teach students how to engage effectively and productively in democratic deliberations”. Within the field of future studies, [4] tries “to depict how the state of inverted totalitarianism is emerging in post-postnormal times, and to illustrate how it shares many of the same features of the totalitarianism depicted in the novels Brave New World (A. Huxley) and 1984 (G. Orwell)”. In the field of semiotics, [5] states that the interesting aspect encountered in Orwell’s 1984 is the vicious unended cycle and the war that will never end between the stated groups within the framework of the ideology/axiology perspective”. When it comes to word embeddings applied to literary arts, [6] extracted a social network from a literary text and “the experiments suggest that specific types of word embeddings like word2vec are well-suited for the task at hand and the specific circumstances of literary fiction text”. One paper [7] demonstrates that building word embeddings on annotated literary texts of 19th century fiction “can provide us with an insight into how characters group differently under different conditions, allowing us to make comparisons across different novels and authors. These results suggest that word embeddings can potentially provide a useful tool in supporting a quantitative literary analysis”. A research conducted by [8] on applying word embeddings and semantic lexicons to literary texts “demonstrates the importance of examining implicit assumptions around default strategies, when using embeddings with literary texts, and highlights the potential of quantitative analysis to inform critical analysis”. Furthermore, projects such as the LAPPS Grid have developed a platform – [9] states that “providing access to a vast array of language processing tools and resources for the purposes of research and development in natural language processing (NLP), has recently expanded to enhance its usability by non-technical users, such as those in the Digital Humanities community (DH)”, but “it is only recently that Computational Linguistics (CL) methods and tools have begun to be made more accessible to non-technical users, and are beginning to be widely adopted by the DH community; however, there remains considerable work to be done to fully adapt CL tools and methods in order to be used by DH scholars”. One paper [10] tried to demonstrate, through the use of a computational analysis of frequencies of dystopian terminology in the text of George Orwell’s 1984, that it is possible to measure how the Orwellian concept is created, constructed and structured in the novel. Such a computational approach, when made with word embeddings, can help to create a “semantic network” of words from Orwell’s novel. Another research [11] showed how conducting a computational concordance analysis of Orwell’s fictional literary work could be applied for the purpose of analysing terms related to the core Orwellian concept. To analyse specific textual corpora, statistical approaches have been proposed [12, 13], with special emphasis on concordances [14, 15], terminology extraction [16, 17] and language modelling [18].

IV. Research

This section is divided into two subsections – Data Set Acquisition and Preprocessing, and Word Embeddings and Vector Representations. The first subsection deals with the process of acquiring the experimental data set, and the following preprocessing phase. The second subsection discusses the applied research approach and selected natural language processing method.

The authors of this paper decided to employ an objective, precise and cost-effective mathematical way to analyse the chosen data set by constructing a vector space model in order to identify semantically similar words in the data set. This paper concentrates on exploring the possibilities of utilising word embeddings in form of word vector representations for the purpose of analysing a very specific data set – a dystopian novel written by George Orwell. This paper analyses not only the quantitate aspects of the resulting word vectors, but also evaluates the results on a qualitative level. The foundations of this research originate in the studies of artificial intelligence and, mainly, natural language processing.

Both studies and their corresponding techniques are applicable in higher education and for purposes of academic curricula – especially in selected courses in information and communication sciences, and computer science, that deal, among others, with various aspects of natural language understanding and generation, computational language analyses, knowledge and information extraction, intelligent machine behaviour etc.

A. Data Set Acquisition and Preprocessing

The experimental data set contained the whole content of George Orwell’s well-known novel 1984, which came out in 1949. No other data was added to the data set.

The already preprocessed version of the data set was initially used for the purposes of [10, 11]. The preprocessing step encompassed various tasks – conversion of file format from HTML to plain TXT with UTF-8 character encoding, getting rid of style and text formatting etc. Superfluous and undesired characters were deleted from the data set using regexes. The data set content was then split into sentences or segments (text
lines not ending with a sentence delimiter), tokenised afterwards using custom tokenisation rules, and manually inspected for any potential errors – e.g. orthographic mistakes, various oversights, missing or faulty characters, such as apostrophes, quotation marks, dashes, hyphens, brackets and other punctuation marks and typographic symbols.

Most of the preprocessing was made with the use of Python, Perl, sed and awk. Once the data set was ready, it was used as the input for generating distributed word vector representations.

B. Word Embeddings and Vector Representations

The vector space model [19] represents an algebraic model for the simplified representation of text or its constituent parts in form of vectors.

Word embeddings are a language modelling and feature learning technique. They allow transforming, i.e. mapping of distinct words from a text (so-called vocabulary) into numbers – generally, real numbers [20-22]. This is needed since machine learning algorithms usually rely on numeric values. Namely, such algorithms are limited by the type of qualified or desired input, and usually cannot work with plain text straight away. Using input that is represented by a corresponding vector of continuous values in a predefined vector space is only one of the possible approaches to this problem in natural language processing. But this way the complexity of the text can be reduced to a purely mathematical problem. Not only that this method reduces the dimensionality of the problem [22], it allows to inspect the contextual similarity of vectors [20] – namely, the word context serves as the principal feature in the word representation model. In very low-dimensional vector spaces, the values in a vector can be interpreted right away by humans.

The classical bag-of-words (BOW) [23] approach typically generates enormous and sparse vectors for a textual input, due to the one-hot word representation strategy. Here the vectors’ dimensions are determined by the vocabulary size of the textual input, as each dimension corresponds to a separate word. If a word appears in a textual input, the matching value in the resulting vector will be a non-zero value. Besides computational complexity as a result of the large number of dimensions, with BOW there is always a risk of model overfitting, a common problem in machine learning – when a model’s prediction performance is very good in a small number of specific situations but terrible in most other cases due to the large amounts of noise that sparse vectors hold. Also, here the vectors are generally unaware of the underlying similarities between words [23].

With word embeddings machines can learn how to create word vector representations, i.e. how to point each word to one specific vector with values that can be inferred from the ways of using individual words. They can preserve the contextual surroundings of a word and, hence, its contextual similarity [20, 21, 23]. Namely, words that appear nearby in a text will also be in close proximity in a vector space [23]. Put differently, words that come from the same or similar context and are utilised in a comparable way can be associated with each other, since they share a similar vector representation, which encapsulates and explicitly encodes their semantic or syntactic similarity, many linguistic regularities, analogies and patterns, or the mutual interrelations of words [24]. Likewise, words that do not share any similarities tend to not share the same context.

That words with similar context share similar meaning is, in fact, not a novel finding. According to the well-known distributional semantics hypothesis in linguistics, “words that are used and occur in the same contexts tend to purport similar meanings” [25]. According to [26], “a word is characterised by the company it keeps”. In other words, the ways of using a particular word define its distinct meaning.

Word vectors are low-dimensional dense vectors of fixed length which makes them computationally very efficient due to the low space and time complexity [23]. The fixed length arises from the fixed and limited vocabulary size of a textual input. Each word is illustrated by a relatively small vector with, usually, hundreds of dimensions. This is much lower than thousands and, possibly, millions (or more) dimensions that are needed for sparse word representations like in the BOW model. A huge benefit of the dense representations is generalisation power [27].

Since word vectors are just a numerical representation of contextual similarities between words, they can be treated and mathematically processed just like regular mathematical vectors. For instance, one could measure the vector similarity by calculating the cosine angle between two non-zero vectors, which is called cosine similarity [23]. The cosine of 0° equals to 1, and it is less than 1 for any angle in the interval (0, π] radians. So, no vector similarity (similarity of 0) is present at a 90° angle. Total similarity (similarity of 1) is expressed as a 0° angle – those are vectors with the same orientation and complete overlap. Two vectors that are diametrically opposed have a similarity of -1.

In natural language processing, one popular algorithm for generating high-quality word embeddings is word2vec [24]. Word2vec is, in fact, a neural network that is capable of recognising similarities within text and automatically producing a set of vectors [19, 20]. But it is not a deep neural network – it can only build numerical representations that can be interpreted by deep neural networks [20].

The word2vec algorithm uses statistics and linear algebra (calculating dot products in co-occurrence matrices, matrix transposition etc.) to learn word vector representations from textual input [19, 20, 22, 24]. It is able to capture the syntactic and semantic coherences in a language. Each relationship is characterised by a relation-specific vector offset, which allows vector-oriented reasoning based on the offsets between words [28]. If sufficient data and context are provided, word2vec can predict the meaning and associations of words quite accurately.

There are two different models within word2vec used for learning word embeddings – Continuous Bag-of-Words model (CBOW) and Continuous Skip-Gram Model.
(CSGM) [19]. CBOW modelling means learning word embeddings for predicting an individual word by observing the word’s adjacent context, whereas CSGM works the other way around – for a provided individual word the corresponding neighbouring words (context) are predicted [24]. Both models base their findings on the local context window, which, as a parameter in the model, can be adjusted to the user’s specific needs.

The context window size impacts the vectors similarities strongly. Large context windows generate vectors that reflect topical similarities, whereas smaller window context windows tend to produce vectors that exhibit more functional and syntactic similarities [27].

In this research, all of the experiments were done on a multicore machine with a 64-bit operating system, an Intel Core i7 processor (4 cores, 8 threads) and 16 GB of RAM. All available threads were used to train the model in order to decrease training time. The authors constructed a 100-dimensional vector space model, which means that for every word present in the model there were 100 features available. The authors used Python 3, Gensim [29] and word2vec for that purpose. Gensim (short for “generate similar”) is a robust open-source vector space modelling and topic modelling toolkit implemented in Python. It was specifically designed to handle large text collections using data streaming and efficient incremental algorithms.

The basis for constructing the model was the preprocessed data set containing George Orwell’s novel 1984. The data set was also lowercased in order to avoid multiple versions of the same word. As for the training algorithm, CBOW was chosen. Window context size, i.e. the maximum distance between the current and predicted word within a sentence, was set to 5, meaning that at the same time 5 words were taken into consideration. Words with a total frequency less than 3 were ignored (minimal count). Generating word vector representations has shown to be very CPU-intensive.

Then for every single keyword, chosen freely by the authors, 20 semantically most similar words (identified though word vector representations) from the same data set were extracted. Similarity was obtained by computing the cosine similarity between a simple mean of the projection weight vectors of the given words and the vectors for each word in the model [29]. Afterwards, the authors carried out a detailed manual qualitative analysis on 6 keywords and, in total, 120 words that were unveiled (minimal count). Those 6 keywords were performed a manual quality analysis on a sample of 6 word vector representations (15%). Those 6 keywords were chosen arbitrarily by the authors, and 5 of them were related to the Newspeak dictionary and the Orwellian concept: “newspeak” – is the fictional language in Orwell’s novel 1984; it is “politically correct” speech taken to its maximum extent; “doublethink” – refers to reality control; “winston” – refers to Winston Smith, the main protagonist that the reader most identifies with in Orwell’s novel 1984; “telescreen” – refers to a fictional surveillance and communication device, which is operated by the ruling Party in a totalitarian system, in order to keep its citizens under permanent observation; “julia” – refers to the fictional character Julia, who pretends to be supporting Big Brother and the ruling Party, but in fact, despises the system; and “war”.

Fig. 1 shows the 100-dimensional vector representation of the word “newspeak”. Since the numbers in the vector are cosine values they can range from -1 to 1.

A. Results of the Experiment

At first, the authors chose to obtain word vector representations for 40 keywords in total – 24 were related to the Newspeak dictionary and the Orwellian concept [30], whereas 16 belonged to the general domain. Keywords related to the Newspeak dictionary and the Orwellian concept were: “newspeak”, “brother” (refers to “big brother”), “thoughtcrime”, “101” (refers to “room 101”), “winston”, “party” (refers to “inner party”, “outer party”, “party members”), “telescreen”, “julia”, “oceania”, “eurasia”, “ingsoc”, “goldstone”, “eastasia”, “doublethink”, “brotherhood”, “vaporized”, “oldspeak”, “crimestop”, “crimethink”, “minipax”, “minitrue”, “duckspeak”, “prole” and “speakwrite”.

Keywords from the general domain were: “war”, “moment”, “people”, “world”, “voice”, “word”, “free”, “freedom”, “time”, “hate”, “love”, “youth”, “best”, “worst”, “good” and “bad”.

From this base of 40 keywords the authors decided to perform a manual quality analysis on a sample of 6 word vector representations (15%). Those 6 keywords were chosen arbitrarily by the authors, and 5 of them were related to the Newspeak dictionary and the Orwellian concept: “newspeak” – is the fictional language in Orwell’s novel 1984; it is “politically correct” speech taken to its maximum extent; “doublethink” – refers to reality control; “winston” – refers to Winston Smith, the main protagonist that the reader most identifies with in Orwell’s novel 1984; “telescreen” – refers to a fictional surveillance and communication device, which is operated by the ruling Party in a totalitarian system, in order to keep its citizens under permanent observation; “julia” – refers to the fictional character Julia, who pretends to be supporting Big Brother and the ruling Party, but in fact, despises the system; and “war”.

V. RESULTS AND DISCUSSION

Here the authors present and discuss the results of the experiment, as well as its implications. Additionally, the authors point out the various downsides and limitations of the experiment in one of the following subsections.

The results of this research have shown some interesting correlations between the selected words. The word “newspeak” can semantically be associated mostly with the word “power”. Newspeak, the fictional language used by the characters of the novel, is the official language of Oceania. It is based on standard English, with words describing “unorthodox” political ideas removed. It is the principal means of the totalitarian ruling Party to retain power, through the control of language. In this context, the correlation between “newspeak” and “power”
is very precise. Fig. 2 shows the top 20 words that are most similar to the word “newspeak”. All words are ranked according to semantic similarity (vector similarity).

\[ \text{newspeak} \]

\[
\{('power', 0.9999126029968263), ('place', 0.999903834090203),
('close', 0.9999024401214679), ('after', 0.999807538364679),
('tele', 0.999761695544428), ('every', 0.999761526847456),
('who', 0.9997891739731420), ('still', 0.999784505992026),
('except', 0.999782601603490), ('being', 0.999773030262609),
('people', 0.9997645616331372), ('great', 0.999764525920201),
('upon', 0.9997645243168592025), ('somewhere', 0.999761343022313),
('worth', 0.9997608463053641), ('joke', 0.999760406955716),
('life', 0.99975360406471), ('war', 0.999754510046149),
('use', 0.999754676720105), ('house', 0.9997521202527051)
\]

Figure 2. Words most similar to the word “newspeak” (according to vector similarity).

The word “doublethink” is semantically connected mainly with “proles”. Doublethink is the power to hold two completely contradictory beliefs in one’s mind simultaneously, and to accept both of them. “Proles” are proletarians, ca. 85% of Oceania’s population. Although not as closely and rigidly observed as members of the Party, proles, as it is stated in the novel, were taught to be inferior through the principles of doublethink.

The word “winston” predominantly corresponds to the word “julia”. Winston Smith and Julia are the two protagonists of the novel and tragic lovers. Their relationship is precisely shown on the semantic level.

The word “telescreen” is primarily related with the adjective “white”, but amongst others, also with the words “table”, “floor”, “street”, “front”, “bed”, “down”, “middle”, “out” and “body” giving a sense that the telescreen, a two-way television and tool of control by the Party, is always unavoidably present.

The word “julia” is mostly linked with the word “it”, but second by similarity is the word “winston”, which confirms their relationship also on the semantic level, as with the previous example of the word “winston”.

The word “war” is most closely affiliated with the word “newspeak”. The Orwellian concept of an ongoing war correlates well with that of the Newspeak language, as the two concepts are the two means of control by the Party, and this is also confirmed on the semantic level.

B. Downsides of the Experiment

The disadvantages and shortcomings of the experiment are highlighted in this subsection. As the authors chose to use all of the unoccupied CPU power (8 threads) in order to increase the training speed, Python’s seed() functionality was, unfortunately, not available due to the limitations of Gensim and controlling hash randomisation [29], and the problems with thread scheduling in common operating systems. Seed is essential for the initialisation of the random number generator and, hence, for the initialisation of vectors for each word with a hash of the concatenation of a given word and the seed value [29]. Using seed() would allow to exactly and fully deterministically reproduce all of the experiments with the vector space model, which is not the case with this research.

Nevertheless, the authors undertook many training runs, and the differences in the resulting vectors were not significant. Only slight modifications were observed, mostly changes in word ranking within vectors – but, semantic similarity, when seen in general, was not altered considerably.

VI. Future Research

The authors plan to significantly increase the dimensionality of the word vectors and analyse the impact on the word vector representations, in form of a cost-benefit analysis (consumption of CPU power, model training time and performance, memory requirements etc.). This should result in additional generalisation power.

Furthermore, the plan is to increment the data set with additional dystopian texts (novels) in order to increase the data set size. It would also be interesting to inspect what exact words contribute positively, and what negatively to the overall word vector similarities. The authors also plan to increase the context size (context window), and the minimal number of word occurrences that should be considered during the process of generating word vectors.

Furthermore, experiments with the Continuous Skip-Gram Model (CSGM) and evaluating the performance differences between the hierarchical softmax and negative sampling approaches should be carried out [24]. The possibilities of the related doc2vec algorithm should also be investigated [23]. Algorithms for vector visualisation, such as the t-Distributed Stochastic Neighbor Embedding (t-SNE) [31], should be tested as well. Alternative models, such as GloVe (Global Vectors for Word Representation) and Facebook’s fastText should also be applied on this data set. GloVe trains on global word-word co-occurrence counts and makes efficient use of statistics [32], whereas the main improvement of fastText over the original word2vec implementation of vectors is the inclusion of character n-grams [21], which allows computing word representations for words that did not appear in the training data (OOV, “out-of-vocabulary” words).

The authors would also like to examine the seed functionality with regard to the cost of training speed and make the generated models freely available online in order to enable experiment reproducibility. Source code could possibly also be rewritten to allow experimenting without lowering the demand on the machine; scoping, with lemmatisation, and implementing stop words and lexicons of function words for filtering purposes.

VII. Conclusion

Word vectors representations are just one of the many possible approaches to computationally representing words. In order to analyse George Orwell’s novel 1984 on a semantic level, the authors chose to build a 100-dimensional vector representation model from the novel’s content, which preserved the relations between the words and its contextual similarities. This approach was chosen due to the fact that words with similar meaning often have similar word embeddings in form of word vectors. This method has shown to have the potential to be applied for studies of literature, in academic curricula and other fields of higher education. For instance, students could investigate a literary work and gain valuable knowledge about it only through the observation of semantically most
similar words identified through word vectors, and without using any other external resources. Students could then, without having any prior knowledge of the literary work, study the internal connections, the work’s narrative (time and space), language and style, the context and meaning of individual words or (fictional) characters within a literary work and, therefore, try to summarise its key concepts on a semantic level.

REFERENCES


