Behind the Dystopian Sentiment: a Sentiment Analysis of George Orwell’s 1984

I. Dunder* 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 - Recent daily political events returned the focus of the general public to a fictional work of art that may yet prove to be one of most influential through the use of its concepts and language – George Orwell’s 1984. The language of Orwell’s novel, specifically its terminology, has been proven by linguists and other researchers to be in common use in media speech and writing. The author’s intent, in which he unanimously succeeded, was to create a dystopian world reminiscent of totalitarian regimes, through the careful use of style and language. Could this intent be somehow investigated and measured by using precise natural language processing methods? Basing this research on related work, the aim of this paper is to evaluate if it is possible to observe how the dystopian atmosphere of the Orwellian concept is created in the novel by applying sentiment analysis on the sentence level, and by exploiting a polarity lexicon as well. Moreover, can a computer by utilising a classification technique based on the Bayes’ theorem demonstrate Orwell’s intention of creating a dystopian atmosphere? In addition, this paper tries to highlight the possible applications of sentiment analysis of fictional work of art in academic curricula.

Keywords - natural language processing (NLP); sentiment analysis; naïve Bayes classifier; George Orwell’s 1984; dystopia; Newspeak; automatic classification; machine learning; information and communication sciences

I. INTRODUCTION AND MOTIVATION

George Orwell’s 1984, one of the most well-known works of English literature, shows to the present day that it may be one of the most influential through the use of its dystopian concepts, known as the Orwellian concept, and its language – specifically, the fictional language “Newspeak” used by the characters in the novel.

Linguists and other researchers have proven the novel’s terminology to be in common use in media speech and writing, even more so in today’s age of fake news.

Orwell’s intent was that of creating a dystopian world reminiscent of totalitarian regimes. This paper tries to show if this intent, created through the careful use of style and language, could be investigated and measured by precise natural language processing methods.

In a previous research on Orwell’s 1984 [1], the authors tried to show, through computational concordance analysis of fictional literary work, how “terms related to the Orwellian concept, such as “Big Brother” or “Newspeak”, are applied in an affirmative and positive way, and that a computer is not suitable for distinguishing the totalitarian implications of those terms, as computers are inherently not able to put them in the context that is in fact a dystopian work of fiction”.

Furthermore, in another paper the authors conducted a computational analysis of frequencies of the dystopian terminology used in Orwell’s novel [2].

Basing this particular research on previous work, the aim of this paper is to investigate and measure Orwell’s intent. By applying sentiment analysis on the sentence level, and by exploiting a polarity data set as well, an evaluation will show if it is possible to observe how the dystopian atmosphere of the Orwellian concept is created in the novel. Moreover, this paper investigates if a computer could demonstrate Orwell’s intention of creating a dystopian atmosphere by utilising a classification technique based on the Bayes’ theorem.

II. MOTIVATION

One differential study [3] with informal and formal writing styles has shown that sentiment analysis has a wide range of applications, from social media texts to poetry. The study accentuated that “literary arts (poems, novels, essays, plays etc.) are the one of the complex examples of formal text”.

On this note, the goal of this paper is to study through computer-based sentiment analysis, if and how the hypothesised and expected atmosphere of “doom and gloom” in a dystopian work of fiction is constructed. As a proof of concept for further research of interesting phenomena, this method could perhaps be used to reveal the implicit rules of the fictional Newspeak language used in the novel. If a novel in general is indeed a complex example of formal text, a novel like Orwell’s 1984, with its specific terminology and a whole language invented for its purposes, is a proof of concept.

This paper also suggests possible application of time-efficient techniques and precise methods in dystopia research and research of literary corpora, as part of e.g. literature studies or linguistics in higher education.

The following related work section shows previous applications of sentiment analysis on literature and other types of texts (and speech).
III. RELATED WORK

This section lists the wide applications of sentiment analysis on literary works of fiction, with an emphasis of its use in higher education and the academic curricula.

One author uses sentiment analysis to “identify concepts that orient us toward a history of emotions in comparative literature” on examples of Spanish literature [4].

Its application varies from the use as a tool for automatic irony detection [5] to the sentiment analysis of English literature using a task-specific semantic ontology [6].


Furthermore, [8] hypothesised that a “sentiment network can be used to distinguish a document’s genre (tragedies versus comedies), detect a given character’s enemies and allies, and model the overall emotional development of a play”.

There have also been studies on word occurrence statistics and emotions in social media, especially Twitter [9, 10, 11]. Moreover, a study evaluated the different techniques and possible applications of emotion recognition in speech [12].

IV. RESEARCH

The following subsections describe the chosen and analysed data set, the applied research methodology and natural language processing techniques.

A mathematical approach was used in order to detect the sentiment in the data set at the sentence level. The research focuses entirely on studying affective states that are present in sentences from the data set. They are examined quantitatively at first, and then evaluated qualitatively by a human evaluator (manually).

All of the research steps derive from the field of natural language processing (NLP). They can be employed in higher education and for purposes of academic curricula, e.g. for efficient data analysis training, teaching information extraction from specific corpora and subsequent analyses etc.

A. Data Set Preparation

The authors of this paper analysed the novel 1984, also known as “Nineteen Eighty-Four”, written by George Orwell and published in 1949.

The whole novel content was used as the testing data set (test set) in this experiment. The authors used a preprocessed version of the data set [1, 2]. Preprocessing of the initially messy data set included file format conversion, saving the data set as a raw UTF-8 plain text file, stripping of any formatting etc. Regular expressions were applied in order to remove redundant characters and to fix specific characters – e.g. 3200 apostrophes were restored.

The data set was split into sentences (or smaller segments where applicable) and then tokenised with specific tokenisation rules. In the final step, the data set was manually checked for errors (e.g. omission of apostrophes, incorrect quotation marks etc.) and corrected.

Python and Perl were used to carry out most of the preprocessing steps. After finishing the process of data set preparation, sentiment analysis was performed.

B. Sentiment Analysis

The main goal of sentiment analysis is detecting, i.e. identifying the subjective opinion, appraisal, attitude or emotion (feeling) in a given text (or speech) with regards to an arbitrary topic or subject [13, 14]. It is an important natural language processing task used nowadays for analysing markets, public relations, politics, product reviews, feedbacks on customer satisfaction and so on.

There can be many possible outcomes of different sentiment analysis approaches. One could, e.g. examine the level of objectivity or subjectivity in a text, or look for the subjects, objects and entities in a text, or try to identify the topic of a text, or just analyse the polarity, i.e. the general sentiment of a text (or its constituent parts) [13, 14].

Very often researchers analyse only the polarity of text – i.e. whether a text is considered positive, negative or neutral – or the subjectivity of a statement in text – i.e. whether a text chunk is considered a subjective opinion or objective fact. Polarity is usually presented as a decimal number in the range of [-1, 1], where -1 and 1 denote negative and positive sentiment, respectively. On the other hand, subjectivity is typically presented as a decimal number in the range of [0, 1], where 0.0 means “very objective” and 1.0 means “very subjective” [13, 15].

Many times, researchers are interested in classifying text according to some external resources (mostly lexicons), which can be crafted for general purposes or task-specific domains.

Text classification is a common semantic-related objective in natural language processing (NLP), with vast practical applications besides sentiment analysis. It has been heavily utilised for spam identification, data filtering and organising, context mining, information extraction, topic modelling, (social) media discourse analysis, business intelligence for making key decisions, trend monitoring etc. The intent of text classification is to classify source text into meaningful systematised, but fixed categories, marked with a discrete classification label [13, 14].

In machine learning, the naïve Bayes classifier is a probabilistic classifier – an algorithm used for learning models and predicting model output, with strong naive assumptions of independence between model features, i.e. predictors [15-17]. It is based on the Bayes’ theorem, as shown in (1).

\[
P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}
\]
The Bayes’ theorem allows one to find the probability of an event \( A \) happening, given that event \( B \) has occurred. \( B \) is called “evidence” and \( A \) is called “hypothesis”. In this case, \( P(A|B) \) is the probability \( P(label|features) \), \( P(B|A) \) is the \( P(features|label) \), whereas \( P(A) \) is \( P(label) \) and \( P(B) \) is the \( P(features) \) probability [17].

Instead of computing \( P(features) \) explicitly, the algorithm calculates the numerator for each label, and normalises them so their sum is 1 [17]. If the classifier encounters an input with a feature that has never been seen with any classification label, then rather than assigning a probability of 0 to all labels (too harsh penalisation), it will ignore that feature [17].

The assumption, however, is that the features are independent of each other. In other words, one feature does not affect the other – that is why the classification procedure is actually regarded as “naïve”. Moreover, all features are considered to have an equal effect on the overall probabilistic outcome, and they jointly determine which classification label should be assigned to a specific input [17].

In order to define the “correct” classification label for a specified input, the classifier calculates the prior probability of each label, which is determined by examining the frequency of each label in the training data set. Then, each feature’s contribution is combined with this prior probability, to obtain a likelihood estimate for each label. The label with the highest likelihood estimate is subsequently assigned to the input value [17].

Classification essentially means choosing the correct label for a given input. In its basic form, classification inspects each input in isolation from all other inputs, whereas the set of labels is defined in advance [17].

Classifying text from a natural language is a process usually comprised of two fundamental steps – training and prediction [17].

During training, a feature extractor is used to convert each textual input into a feature set – this is done as part of the so-called feature extraction process. The feature set is typically a numerical representation of the provided text in form of a feature vector. Each element of the feature vector symbolises the frequency of a word in the training data set.

These feature sets hold the basic information about each input that should be used to classify it. Then, pairs of feature sets (feature vectors) and classification labels are fed into the machine learning algorithm in order to create a classification model. Basically, the model learns how to associate a specific input, in this case, the text from Orwell’s novel, to the correct output (labels “pos” or “neg”) with regard to the content from the training data set.

During prediction, the same feature extractor is employed to convert unseen textual inputs into feature sets (feature vectors). These feature sets are then passed to the classification model, which generates predicted classification labels.

Here the authors of this paper chose the classical bag-of-words (BOW) approach, which records words along with their frequency. At first, a vocabulary is created by collecting all different words that occur in the training data set, and each word is associated with a count of how it occurs [16].

The BOW model is a simplification, which greatly reduces the level of language complexity, and allows a sentence to be represented as a plain “bag” with a mixture of its inherent words, but at the cost of losing some valuable information about the surroundings of the word. Nevertheless, for sentiment analysis the end result is mainly determined by a few high-impact keywords that provide critical semantic information.

This “bag” is typically implemented as a multiset – a set modification that allows for multiple instances of each of its elements. BOW does not take any grammar, syntactic knowledge or word order into account, but just keeps track of word occurrences and multiplicity, i.e. the number of times a multiset member appears in a multiset [18].

Many recent sentiment classifiers make use of sentiment lexicons [13, 15], which act as a compact representative of emotions from the real world. Those lexicons are in fact lists of words and corresponding sentiments they convey. However, as they can be of different types and domains, crafting lexicons is a very distinct process and varies from task to task. So it is not surprising that there is no gold standard, i.e. general agreement on how to treat and label the very subjective affective states within individual lexicons.

In this experiment, the authors chose to approach the sentiment analysis task as a typical classification problem in natural language processing. This was done computationally by employing an automatic classification method.

For that purpose, a machine learning model was trained in form of a classifier, which was then used to discriminate words based on their frequency, i.e. the number of occurrences of each word. The frequency was, thus, used as the training feature in the model. The main idea was to forward text, i.e. sentences from Orwell’s novel, to the classifier and obtain the corresponding classification labels.

Here, the authors used supervised machine learning and single-closed-class classification:

- classification was “supervised” since the classifier was generated using a training data set (polarity lexicon) which contained the correct label for each input – the machine learning algorithm learned during the training phase what the expected output was for a given input sample;
- each instance could be assigned only to a single label at a time;
- all the possible labels were determined beforehand (closed class).

The authors had to process the data set at first, then a feature extractor was applied in order to extract the features (which words in the training data set occur in the
data set). The machine learning algorithm learned these features in order to predict against unseen, i.e. new text – in this case, Orwell’s novel 1984.

The authors created a naïve Bayes classifier by passing the training data set in form of an already available polarity lexicon [19]. The lexicon consisted of 6790 words in total. All words were labelled either with the label “pos” (for positive sentiment) or “neg” (for negative sentiment):

- 4783 words labelled with “neg” (ca. 70.4%);
- 2007 words labelled with “pos” (ca. 29.6%).

Neutrally labelled words were not present in the training data set, i.e. polarity lexicon. Although the lexicon was decent in size, machine learning typically relies on larger high-quality training data sets, due to the fact that the classification model depends on the quality and quantity of features (predictors). The lexicon was converted into CSV file format for easier later-stage processing and analyses.

In this experiment, the implementation of the naïve Bayes classifier in Python was fairly simple, but the model itself had a significant downside. As already mentioned before, features should be completely independent, but in real natural languages, and especially in literature, this is very often not the case. Certainly, this obstructed the classification accuracy.

V. RESULTS AND DISCUSSION

This section deals with the experiment results and discusses the various experimental aspects. In addition, the authors list the different drawbacks of the conducted experiment in the second subsection.

A. Experiment Results

The experiments were carried out on a 64-bit quad-core machine (8 threads) with 16 GB of RAM. Training the classifier took relatively long. More than 5.2 GB of RAM were consumed during the training phase due to the reasonable size of the training data set.

All of the 6905 sentences (or lines in same cases) from the data set, i.e. from George Orwell’s novel 1984, were automatically classified as either positive (“pos”) or negative (“neg”). Lines were understood as text fragments present in the data set that were not ending with a punctuation mark, i.e. with a sentence delimiter.

The authors decided to perform a manual quality analysis of the trained classifier afterwards. For that purpose, 300 sentences/lines were randomly extracted from the original data set (about 4.3% of the entire data set). The number 300 was chosen arbitrarily by the authors. Out of the extracted subset, 204 sentences/lines were classified as “neg” by the classifier, and 96 as “pos”.

One evaluator checked if the classifier had correctly classified the 300 sentences/lines, and to what extent – in total, 134 were manually marked as “correctly classified”, whereas 165 were marked as “incorrectly classified”. In other words, only ca. 45% of all the 300 randomly extracted and qualitatively evaluated sentences/lines were confirmed to be accurately classified. On the other hand, 55% were erroneously classified, which indicates that the trained classifier performed poorly. Some of the reasons for the insufficient quality will be discussed in the next subsection.

But by and large, in order to provide more adequate conclusions on whether this applied natural language processing technique is suitable for analysing dystopian novels, or other related literary works, more extensive research is needed and planned for future work.

B. Experiment Drawbacks

The authors of this paper would also like to emphasise the caveats and limitations of the experimental setup.

During the manual evaluation of the 300 randomly selected sentences, on some occasions it was not easily determinable whether a sentence was correctly classified, due to a number of reasons:

- the only two labelling options were “pos” (positive) and “neg” (negative), since there were no words present in the training data set (polarity lexicon) that were labelled as “neutral”, which would sometimes definitely be much more suitable;
- the training data set (polarity lexicon) was of general domain, and did not contain any of the specific Newspeak words (fictional language very frequently used in Orwell’s 1984), which made the correct classification sometimes impossible;
- analyses of expressed affective states were occasionally challenging and difficult due to the linguistic and stylistic complexities of the selected novel – such as sarcasm, metaphors, irony, specific terminology etc.;
- in some of the sentences the classification labels were hard to differentiate due to the tone and nuance of particular expressed subjective opinions;
- observing sentences as isolated units and, hence, the lack of larger context, made it oftentimes hard to examine the correctness of the classification;
- the semantics of some of the sentences could not be accurately interpreted, as affective states were not clearly and explicitly stated, but the semantic information derived from other sentences located elsewhere in the novel;
- the lack of directly expressed opinions made it sometimes hard to grasp what the sentence was truly about;
- analysing some of the sentences was very demanding, as the necessary word span for correctly identifying different semantic concepts was very large (e.g. when one sentence refers to another, large sentence
Also, whenever the evaluator came across a dubious sentence or dilemma, and was not sure about the correct classification, the preference was given to the “neg” label. Also, if the classifier labelled a completely neutral sentence/line as “pos” or “neg” the evaluator marked such an occasion as a classification error. This was indeed crucial and resulted in additional, but artificial, penalisation (decreasing) of the classifier’s accuracy during the manual evaluation phase.

Furthermore, the uneven distribution, i.e. disproportion in the number of negatively and positively labelled words in the polarity lexicon skewed the probability results towards an overall negative sentiment. There was just a higher chance of bumping into a negatively labelled word. Namely, almost 2.5 times more “neg” labels were present in the polarity lexicon, which was used to train the classifier in the first place. However, this did not have a purely bad effect on the evaluation outcome, and was not entirely noticeable, as in Orwell’s work – a dystopian novel – the negative sentiment is dominant anyway.

VI. FUTURE RESEARCH

The authors plan to increase the training data set (polarity lexicon), i.e. to increment it with all of the specific Newspeak words from Orwell’s novel.

Furthermore, a specific domain lexicon should be crafted with specific (dystopian) vocabulary from the 20th century literature in order to capture the zeitgeist of 1940s and 1950s world, i.e. the time frame when Orwell’s novel 1984 was written and published. Using pondering (favouring specific terminology) in the classification model could also be useful.

It should also be explored what other sentiment lexicons exist on the market, and which of them are freely available and editable.

Also, adding the “neutral” category with corresponding neutral words to the lexicon is important in order to allow for a more fine-grained sentiment analysis.

Perhaps adding more polarity labels, i.e. categories (e.g. “very positive”, “positive”, “neutral”, “negative”, “very negative”) would also be beneficial to the classification quality and precision.

Here the sentiment analysis is done only on the level of a single sentence/line which leads to a certain loss of semantic information. Due to the fact that most sentences are mutually not independent and the lack of larger context very often hinder the evaluation task and make an objective discrimination of polarity labels very complicated.

One specific CAT tool [20] provides means for discovering concordances, and showed to be very effective in context analyses. The authors plan to investigate the possibilities of using it for fast lookups of concordances or its fragments.

The authors plan to increase the validity of the manual quality evaluation by conducting sentiment analyses:

- on the level of the whole corpus, i.e. data set – a final sentiment score is given for the entire document (Orwell’s novel);
- on the level of multi-word expressions – which are treated as a single grammatical unit – collocations and idiomatic expressions within a sentence/line [21, 22];
- on the aspect level of specific text chunks within sentences/lines, which can be identified with part of speech (POS) tags.

In this experiment only one evaluator did the qualitative analysis of the classifier’s output. More evaluators should repeat this experiment in order to verify the performance of the classifier. Moreover, an inter-rater agreement analysis could provide interesting insights into the consensus and reliability of human judgement [23], i.e. the difficulties of accurate classification of dystopian novel content by humans.

Besides, a larger evaluation data subset should be used (much more than 300) and, later on, F-measure, as the harmonic mean of precision and recall [24], should be calculated for more elaborate evaluations of the classifier accuracy.

The classifier in this experiment behaved naïve. To be more precise, it actually classified Orwell’s novel on the word level – the sentiment analysis score of an entire sentence/line derives from the word level probability calculations of words that comprise them. That way sentence level classification did not consider any linguistic knowledge or syntactic information on how sentences and paragraphs are structured. The possibilities of applying appropriately trained language models and calculating the perplexity of the model [23] could help to increase the classifier’s predicting capabilities.

Source code should partly be rewritten to speed up training, reduce loading time and to enable quicker data access later on. The authors plan to apply alternative data structures and object serialisation (pickling) for faster loading, which is especially important for consecutive analyses and evaluation runs.

VII. CONCLUSION

Sentiment analysis is fast and unbiased, and, unlike human subjective judgment, applies consistent criteria. It relies purely on a mathematical framework, deprived from any personal points of view, assumptions, impressions, opinions, understanding, feelings, convictions and reasoning.

The authors have shown that the absence of the neutral label in the polarity lexicon makes assessing a classifier’s accuracy more difficult. Still, defining the exact meaning of the “neutral” sentiment is not always straightforward. Even though identifying some expressions as neutral can also be a demanding task, granularity has shown to be very valuable in sentiment analysis.

In addition, resorting to a general domain polarity lexicon did not prove to be beneficial to the quality of classification of dystopian literature.
Also, qualitative manual evaluation and error analysis have demonstrated to be very context-sensitive. The meaning of a given sentence/line is, most of the times, inferred from the surrounding text. That is why during the evaluation phase a larger context window should be presented to the evaluator. This should help to reduce errors and improve evaluation consistency.

The authors are convinced that the applied research methodology can be adopted in academic research and education, and that automatic natural language processing approaches indicate a huge potential, especially since numerous problems and experimental drawbacks have been identified, out of which all can be taken into consideration in future research.

Despite the low accuracy of the classifier, the results still look promising. Further investigation should yield definitive conclusions on whether an automatic natural language processing approach, in form of sentiment analysis, is appropriate for analysing the dystopian literature and its unique style.

REFERENCES


