Good, Neutral or Bad - News Classification

Aashish Agarwal, Ankita Mandal, Matthias Schaffeld, Fangzheng Ji, Jihao Zhang, Yiqi Sun
University of Duisburg-Essen
Duisburg, Germany
{firstName.lastName}@stud.uni-due.de

Ahmet Aker
University of Duisburg-Essen
Duisburg, Germany
a.aker@is.inf.uni-due.de

Abstract

Reading news articles affects the mood and mindset of the reader. Therefore we want to provide means to track our daily news consumption activities. In this paper, we release news articles dataset assigned with good, bad and neutral labels. The dataset comprises of 300 news articles, each annotated by five different annotators. The agreement among the annotators is 0.526 according to Krippendorff's Alpha and 0.435 according to FleissKappa. We also experiment with four different machine learning approaches such as Naive Bayes, SVM, Logistic Regression and Deep Learning using LSTM units. Our experiments show that NaiveBayes significantly outperforms the other three classifiers.

1 Introduction

In the media, the presence of bad news seems to dominate over good news. Every day there is at least a report about terrorism, natural or human-made disaster, a war crime, human right violation, airplane crash, etc. Studies show that news, in general, has a significant impact on our mental stature [8]. However, it is also demonstrated that the influence of bad news is more significant than good news [13, 2] and that due to the natural negativity bias, as described by [11], humans may end up consuming more bad than good news. This is a real threat to the society as according to medical doctors and, psychologists exposure to bad news may have severe and long-lasting negative effects for our well being and lead to stress, anxiety, and depression [8]. Furthermore, specific kinds of bad news, for example about unemployment, may affect stock markets and in turn, the overall economy [4].

In our ever-digitized world, with a constant influx of news from a variety of sources, differentiating good and bad news may help the reader to combat this issue. A system that filters news based on the content of the article, no matter the news website a person is following, may enable the user to control the amount of bad news they are consuming. Whilst most people start their day with reading the news, they can then start it on a positive note.

To implement such a news filtering system we created a gold standard dataset comprising 300 news articles annotated by five different raters with good, bad and neutral labels. This dataset will be made publicly accessible and can be used for further research.[1]

The definitions of good, bad and neutral news may widely vary from individual to individual and from country to country [7]. Therefore, we defined three categories explicitly - what can be termed as good, bad or neutral news. To measure the quality of the ratings we used Fleiss Kappa and Krippendorf's Alpha to check for inter-rater reliability. We also evaluated several machine learning techniques including Naive Bayes, Logistic Regression, Support Vector Machines

In: A. Aker, D. Albakour, A. Barrón-Cedeño, S. Dori-Hacohen, M. Martinez, J. Stray, S. Tippmann (eds.): Proceedings of the NewsIR'19 Workshop at SIGIR, Paris, France, 25-July-2019, published at http://ceur-ws.org

[1] https://github.com/ahmetaker/goodBadNews

and Deep Learning on the collected dataset. These four techniques should give the first impression on the complexity of the task and serve as baselines to further improve the results. Our initial results show that Naive Bayes significantly outperforms the other three approaches.

In the first section of the paper, we define the terms good, bad and neutral news. We also describe the process of corpus collection and agreement on ratings. Next, in Section 3, we describe our methods of feature engineering and our baseline methods. In Section 4 we present our results. Finally, we conclude the paper in Section 5 with what can be done as future work.

2 Corpus

2.1 Definition of good, bad and neutral news

According to the Collins English dictionary[2] good news is defined as "someone or something that is positive, encouraging, uplifting, desirable, or the like" and bad news "someone or something regarded as undesirable". For neutral news, we stated that neither of this is the case. We used these definitions to start our annotation. With these definitions, we run an initial annotation process with 20 randomly selected news articles. We asked 5 annotators who were undergraduate students, with ages varying from 20-25 years, fluent in English and frequent online news readers to read the news and provide good, bad or neutral label according to the above definitions. However, our annotators found these definitions not unambiguous enough so that we revisited the design of our guidelines. This included using an exemplified definition instead. In the following we briefly outline these exemplified definitions:

Good News If the subject of the article is someone being saved from danger, the creation of medicine which can cure or help with an illness, the end of a war or some kind of disaster, human rights being defended, or something that benefits the public or a dangerous culprit being arrested.

Neutral News If the subject of the article is a popularization of science, history or geography, describing humanistic traditions, astronomy, nature, history or landscape, scientific literature, news of people's livelihood without casualties or daily entertainment and fashion news.

Bad News If the subject of the article is a war, accidents, disaster, epidemic disease or killing, criminal activities, the death of a famous or important person, some sort of discrimination, bullying or stereotypes, some negative influence or event regarding economics, nature, animals or human rights.

Number of Articles	300
Average Sentences Count	24.23
Average Word Count	497.83
Number of good news	52
Number of bad news	131
Number of neutral news	117

Table 1: Statistics about the corpus

Fleiss Kappa	0.435
Krippendorffs Alpha	0.526

Table 2: Inter-rater agreement

Using these exemplified definitions we re-run the annotation process with another randomly selected 20 articles and this resulted in more satisfactory annotations so that we used this strategy to create our corpus.

2.2 Corpus Collection

Using Newspaper3k[3], we randomly collected a corpus of 300 English news articles[4]. The articles come from different news agencies such as BBC.co.uk, independent.co.uk and entail topics from categories such as economic, medical, international, local and emergent news. We used the exemplified definitions given above to annotate these as good, bad or neutral news. The same five undergraduate students as above took part in the annotation task. After gathering the annotations for all news articles, we took the majority of the readers' opinions as the final definition. If no clear majority vote was found, we introduced a meta reviewer who was not among the five annotators to give a final decision. Table 1 gives some stats about the corpus as well as the distribution of the different classes.

We also computed the agreement among the annotators. To do this, we used Fleiss' kappa and Krippendorff's alpha. Table 2 shows the results for inter-rater agreement. From the table, we can see that the agreement is moderate indicating the difficulty of the task.

3 Experiment

The task of good, bad or neutral news classification is to classify a given online news article to one of those classes. To find a classifier suited for this task, we explore different traditional machine learning approaches as well as deep learning. In both cases, we only use the article content to extract features. More precisely, for the traditional machine learning techniques we use Bag of Words (outlined in the next Section) and for

[2] https://www.collinsdictionary.com/

[3] https://pypi.org/project/newspaper3k/

[4] These 300 articles are exclusive from those 40 articles used to refine the annotation definitions.

deep learning the lead parts of each article represented with word embeddings.

3.1 Feature Engineering

For the traditional machine learning approaches, we use Bag of Words (BoW) as the only feature category. In total, our vocabulary contains 19000 tokens including stop words, digits, inflected forms of the words, etc. We use the following pre-processing steps to reduce the vocabulary size to 13000 words:

- Lower casing the article texts.
- Removing stop-words.
- Removing digits and punctuation marks.
- Removing contractions.
- Depicting all numbers as #.
- Lemmatizing the words.

Each of the words is represented using term frequency (TF) (number of times a word occurs in a particular news article) and inverse document frequency (IDF) (number of articles from the corpus the word appears in). We further reduce the vocabulary size by only using the significant words. For this, we use the Chi-square test and select those words that were significant in discriminating the classes. After this step, the vocabulary contains around 3600 words. We use these words represented using TF*IDF to guide our traditional machine learning approaches.

For the deep learning technique, we use the lead part of each article, convert each word in this part into word embeddings and use these to represent each article.

3.2 Baselines

As baselines, we experiment with Naive Bayes classifier, Support Vector Machines, Multinomial Logistic Regression and a deep learning model using LSTMs.

Naive Bayes is often used in text classification applications and experiments because of its simplicity and effectiveness [10]. It uses a probabilistic model of text. Naive Bayes classifier is highly scalable, requiring several parameters linear in the number of variables (features/predictors) in a learning problem [12]. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. Determined by grid-search, we set alpha to 0.01.

Logistic Regression is one of the most popular supervised classification algorithms. Multinomial Logistic Regression is the generalization of the Logistic Regression algorithm which can be used to conduct when the dependent variable is nominal with more than two levels. It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables. Using Grid-search, we set C to 50 and regularization to l2.

The ***SVM*** problem is to find the decision hyperplane that can maximize the margin between the data points of the classes [5]. Corresponding to our Grid-search analysis, we use a linear kernel and set C to 10.

Our ***deep learning*** model comprises a simple **LSTM** layer [1] that is capable to consider sequential information. The input of the LSTM (50 LSTMs) layer is word embeddings. We obtain the embeddings from the input documents. Note, as stated above instead using the entire article as input we use only the lead part of each article which can be considered as the summary of news article [14]. For simplicity and also to have a common input length across all the articles we use the first 400 words of each article as the lead part of the article. We use a Dropout layer after the LSTM (0.1), which is followed by a dense layer (50 units with ReLu activation) and then again by a Dropout layer (0.35) and finally by a SoftMax layer. We use *Adam* as the optimization function with *0.001* learning rate and *Xavier Initialization* for weight initialization. The loss is determined by *categorical crossentropy* together with *l2 regularization.* Our batch size is 64, and Epoch number is set to 40.

4 Results

The results of the performances of the different classifiers are presented in Table 3. In all cases, we used 10-fold cross-validation and report in macro-averaged F1 measure, precision and recall. From the results, we see that the best performing classifier is the Naive Bayes outperforming all the other classifiers. Significance test using paired t-test with Bonferroni correction ($p < 0.0125$) [3] shows that the Naive Bases classifier significantly outperforms the other classifiers.

5 Conclusion and Future Work

In this paper, we propose to release a dataset containing news articles annotated with good, bad and neutral labels. We have a total of 300 news articles in our dataset where each article has been annotated

Classifier	Accuracy	Precision	Recall	F1
NaiveBayes	0.829	0.828	0.796	0.799
SVM	0.717	0.517	0.583	0.533
LogReg	0.700	0.475	0.565	0.511
LSTM	0.594	0.415	0.478	0.533

Table 3: Overall Classifier Performance Comparison

by five different annotators. We computed the interrater agreement using Krippendorff's Alpha and Fleiss Kappa. According to Krippendorff's Alpha, the agreement is 0.526 and according to Fleiss Kappa 0.435. We also experiment with four different machine learning approaches such as Naive Bayes, SVM, Logistic Regression and Deep Learning using LSTM to provide initial results on the task. Our experiments show that Naive Bayes significantly outperforms the other three classifiers.

In the future, we plan to extend the dataset. This would allow the approaches to gain more stability, especially the deep learning strategies whose performance rely on bigger training data. We also plan to investigate features other than Bag of Words to capture sentiments, emotions and similar linguistic aspects that better distinguish between bad and good news.

6 Application

Nowadays, the amount of online news content is immense and its sources are very diverse. For the readers and other consumers of online news who value balanced, diverse and reliable information, it is necessary to have access to additional information to evaluate the news articles available to them. For this purpose, Fuhr et al. [6] propose to label every online news article with information nutrition labels to describe the ingredients of the article and thus give the reader a chance to evaluate what she is reading. This concept is analogous to food packages where nutrition labels help buyers in their decision making. The authors discuss 9 different information nutrition including sentiment, subjectivity, objectivity, ease of reading, etc. We propose the bad/good/neutral classification as an additional information nutrition label and plan to implement this in our freely available News-Scan[5] tool [9]. This tool is a browser plugin that can be evoked by users to obtain nutrition labels for the articles they are currently reading.

7 ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media".

[5] www.news-scan.com

References

[1] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. Bad is stronger than good. *Review of General Psychology 5*, 4 (2001), 323–370.

[3] Bland, J. M., and Altman, D. G. Multiple significance tests: the bonferroni method. *Bmj 310*, 6973 (1995), 170.

[4] Boyd, J. H., Hu, J., and Jagannathan, R. The stock market's reaction to unemployment news: Why bad news is usually good for stocks. *Journal of Finance 60*, 2 (2005), 649–672.

[5] Colas F., B. P. Comparison of svm and some older classification algorithms in text classification tasks. in: Bramer m. (eds) artificial intelligence in theory and practice. *IFIP International Federation for Information Processing 217* (2006).

[6] Fuhr, N., Nejdl, W., Peters, I., Stein, B., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., and Mothe, J. An information nutritional label for online documents. *ACM SIGIR Forum 51*, 3 (feb 2018), 46–66.

[7] Giner, B., and Rees, W. On the asymmetric recognition of good and bad news in france, germany and the united kingdom. *Journal of Business Finance & Accounting 28*, 910, 1285–1331.

[8] Johnston, W. M., and Davey, G. C. L. The psychological impact of negative tv news bulletins: The catastrophizing of personal worries. *British Journal of Psychology 88*, 1 (1997), 85–91.

[9] Kevin, V., Högden, B., Schwenger, C., Sahan, A., Madan, N., Aggarwal, P., Bangaru, A., Muradov, F., and Aker, A. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (Brussels, Belgium, Nov. 2018), Association for Computational Linguistics, pp. 28–33.

[10] Kim S. B., Rim H. C., Y. D. S., and S, L. H. Effective methods for improving naive bayes text classifiers. *LNAI 2417* (2002), 414–423.

[11] Rozin, P., and Royzman, E. B. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review 5*, 4 (2001), 296–320.

[12] Russell, Stuart; Norvig, P. *Artificial Intelligence: A Modern Approach(2nd ed.)*. Prentice-Hall, 2003.

[13] Soroka, S. N. Good news and bad news: Asymmetric responses to economic information. *The Journal of Politics 68*, 2 (2006), 372–385.

[14] Wasson, M. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics* (1998), vol. 2.