A Relativistic Opinion Mining Approach to Detect Factual or Opinionated News Sources

Erhan Sezerer[1] and Selma Tekir[2]

[1] Izmir Institute of Technology, 35430, Urla, Izmir, Turkey erhansezerer@iyte.edu.tr
[2] Izmir Institute of Technology, 35430, Urla, Izmir, Turkey selmatekir@iyte.edu.tr

Abstract. The credibility of news cannot be isolated from that of its source. Further, it is mainly associated with a news source's trustworthiness and expertise. In an effort to measure the trustworthiness of a news source, the factor of "is factual or opinionated" must be considered among others.
In this work, we propose an unsupervised probabilistic lexicon-based opinion mining approach to describe a news source as "being factual or opinionated". We get words' positive, negative, and objective scores from a sentiment lexicon and normalize these scores through the use of their cumulative distribution. The idea behind the use of such a statistical approach is inspired from the relativism that each word is evaluated with its difference from the average word. In order to test the effectiveness of the approach, three different news sources are chosen. They are editorials, New York Times articles, and Reuters articles, which differ in their characteristic of being opinionated. Thus, the experimental validation is done by the analysis of variance on these different groups of news. The results prove that our technique can distinguish the news articles from these groups with respect to "being factual or opinionated" in a statistically significant way.

1 Introduction

In today's continuous news flow by various news providers the following two questions have gained importance: What makes a news source credible? What are the most trustworthy news sources?

Journalism literature analyze news source credibility through the dimensions of trustworthiness and expertise [Gaziano and McGrath (1986)]. The dimension of trustworthiness is further examined by the factors "is factual or opinionated", "is biased or unbiased", "does or does not separate fact and opinion", and "does or does not tell the whole story" [Gaziano and McGrath (1986)].

This work covers the factor of "is factual or opinionated" as part of an effort to measure the trustworthiness of news sources. To address this factor, we propose a novel unsupervised probabilistic lexicon-based opinion mining technique. For separating facts from opinions, we take our intuition from the relativism. Our relativistic perspective has two implications; first, it can be said that objectivity is not something that can be computed on its own, rather it is the lack of any positive or negative sentiment. Second, for calculating negativity and positivity we should find how different a word is (how much more negative/positive) from the average word used in English. In order to find this relative strength of the words, we first get the sentiment (positive, negative, and objective) scores from a sentiment lexicon and normalize those scores by their cumulative distribution. Although the proposed technique is classified as a lexicon-based technique, the nature of calculating scores of words relativistically, allows this method to be applied along with any kind of sentiment analysis technique.

We have used three different news sources in our experimental setup: Editorials, New York Times articles, and Reuters articles. Editorials are opinion pieces that are written or evaluated by the editorial staff. The New York Times articles are the regular news articles and Reuters articles are mainly factual wire agency pieces. Therefore, each of these three sources demonstrates different level of opinionatedness. As an experimental hypothesis we claim that editorials are more opinionated than the regular New York Times articles and these regular articles are more opinionated than the articles of the Reuters Agency. Our experimental results prove this claim statistically.

Our contribution comes in two different ways. To our knowledge there isn't any work that examines the credibility of news sources through the use of fact/opinion ratio and second our method is the first method that uses Cumulative Distribution-based normalization in sentiment scores.

In the remaining part of the article, we first refer to existing literature and our contribution in the context. Section 3 describes our experimental setup covering the dataset, knowledge-base and preprocessing, and sentiment analysis parts. In section 4, we present our experimental results. Finally, we conclude our work and state possible future research directions.

2 Related Work

Opinion mining, sentiment analysis, and/or subjectivity analysis deal with the computational treatment of opinion, sentiment, and subjectivity in text [Pang and Lee(2012)]. By opinionated, it means that a document or sentence expresses or implies a positive or negative sentiment [Liu(2012)]. Classifying an opinionated text as either positive or negative is termed as document-level sentiment-polarity classification problem [Pang and Lee(2012)].

The emerging area of contradiction analysis models and analyzes conflicting opinions. Subjectivity analysis is the general term that covers opinion mining, sentiment analysis, opinion aggregation, and contradiction analysis [Tsytsarau and Palpanas(2012)].

Our work is an unsupervised probabilistic lexicon-based opinion mining technique that is applied to news texts to determine their level of being factual or opinionated in an effort to measure the credibility of the providing news source. Our related work is structured as follows: First, the different approaches of opinion mining is described. After that, similar work on news opinion mining is given.

In their survey on subjectivity analysis, Tsytsarau and Palpanas (2012) classifies the different approaches of opinion mining as machine learning and dictionary-based, the latter including the corpus statistics and semantic approaches.

A recent survey on multilingual sentiment analysis [Dashtipour et al.(2016)] goes over the state-of-the-art sentiment analysis techniques and compares them on two standard datasets both in terms of their accuracy and computational cost. According to the accuracy, their comparison states that Support Vector Machines (SVM) is the best supervised method. In the unsupervised category, on the other hand, semantic orientation method of Singh et. al. (2013) outperforms the others.

The machine learning methods assign sentiment scores to documents based on a model learned from document feature vectors and their known sentiments. The document feature vectors are constructed either according to word presence or word frequencies. Pang et al. (2002) in their work “Thumbs up?” predicts the sentence-wise sentiments using the features mainly based on word presence by the classification algorithms of Naive Bayes, Maximum Entropy, and Support Vector Machines (SVM). Their results show that the predicted sentiment scores outperform the random baseline and SVM has the superior performance.

Unsupervised approaches to document-level sentiment analysis are based on determining the semantic orientation (SO) of specific phrases within the document. If the average SO of these phrases is above some predefined threshold the document is classified as positive and otherwise it is deemed negative. There are two main approaches to the selection of phrases: a set of predefined part-of-speech (POS) patterns can be used to select these phrases or a lexicon of sentiment words and phrases can be used [Feldman (2013)].

The dictionary-based methods take the sentiment of each word from a dictionary and aggregate the sentiment values of words through a weighting formula in order to calculate the sentiment value of a text segment. The recent work in this category uses different document sentiment aggregation formulas that make use of direct polarity scores from dictionaries.

The unsupervised semantic orientation method [Singh et al.(2013)] is a state-of-the-art lexicon-based method to classify documents (reviews) as either positive or negative. In doing so, they make use of the semantic orientation values of adjectives. They extract adjectives and aggregate the semantic orientation values of them for the document or review. If the value exceeds a given threshold, the review is tagged as positive, otherwise it’s evaluated as negative.

The concept of news credibility with its subdimensions is introduced first by Gaziano and McGrath (1986). They define the credibility by trustworthiness and expertise of news source. Moreover, they decompose the dimension of trustworthiness into more concrete factors. Among those factors, “does or does not separate fact and opinion”, and “is factual or opinionated” stand out.

Yu and Hatzivassiloglou (2003) uses a Naive Bayes classifier to distinguish articles under News and Business (facts) from articles under Editorial and Letter to the Editor (opinions). Their document-level classifier is evaluated with respect to article type labels (preassigned opinion and fact labels) that are part of the meta-data by the news provider. Our approach differs from this work in three important ways: First, we propose to separate facts from opinions as part of an effort to measure the credibility of news sources. Second, they use a supervised machine learning algorithm to separate facts from opinions at the document level using different features like the average semantic orientation score of the words. We use an unsupervised, statistical technique to make this separation at the corpus level. Last, our experimental design differs. We have three groups of article sets that are known to be different in being factual or opinionated.

The experimental hypothesis is our technique is influential in identifying the degree of being factual or opinionated. Analysis of variance is applied to test this hypothesis.

Morinaga et. al. (2002) and Bethard et. al. (2004) create an opinion-indicator lexicon by looking for terms that tend to be associated more highly with subjective-genre newswire, such as editorials, than with objective-genre newswire [Pang and Lee(2012)].

3 Experimental Setup

In this section we describe our dataset and the details of our lexicon-based sentiment analysis technique.

3.1 Dataset

Our data consists of articles from three different sources: The New York Times' Editorials, The New York Times' articles [Sandhaus(2008)] and Reuters' articles [Rose et al.(2002)]. The New York Times (NYT) corpus contains approximately 1.800.000 articles, published in New York Times between January 1st. 1987 and June 19th. 2007. The Reuters corpus has around 800.000 articles, published between July 20th, 1996 and July 19th, 1997.

The reason behind this dataset selection is, they have a distinguishing property that we will try to verify with our tests. While The New York Times is a news source that publishes regular news articles, Reuters is a news agency, and because of this Reuters articles are expected to be briefer and freer from opinion of the authors. On the other hand, The New York Times, although it can be very objective too, is more likely to have opinions. This assumption comes from the fact that, the aim of Reuters and such wire agencies is to provide information to other news sources, and to do so, it should be, and in most of the cases actually is, more fact-based and pure. As Fenby states [Fenby(1986)], objectivity is the philosophical basis for wire agencies and as these wire services are preselectors and preprocessors of news the emphasis had to be on hard facts rather than comment. Regular news sources, on the other hand, have a particular audience to address, so they tend to have more emotions and opinion on articles to conform with the general view of their audience. Although it is stated that there should be a difference between the two sources in terms of subjectivity, both sources are expected to be objectively written in contrast to the editorials which are written for the purpose of projecting the author's ideas.

By using these three types of data, we aimed to create an environment with three levels of objectivity: editorials, news articles and news agency articles where the objectivity increases from the first to the last.

3.2 Knowledge-base and preprocessing

In our work, for calculating words' score we used SentiWordNet [Esuli and Sebastiani(2006)], which is a sentiment dictionary that relies on the WordNet [Miller(1995)] dictionary for English words. Since the SentiWordNet scores individual words according to their POS tags, we used maximum-entropy (CMM) part-of-speech tagger of Stanford NLP Group [Toutanova et al.(2003)].

3.3 Sentiment Analysis

We have built this method on the idea of relativism by making two important analogies to it. First, it can be said that objectivity is not something that can be computed on its own, rather it is the lack of any positive or negative sentiment. Any kind of calculation on objectivity should be built relatively to the negative/positive scores. So, computing the objectivity scores without considering the negative or positive scores will not help us in achieving our goals. Second, calculating negativity and positivity is an important issue in itself. The analogy to relativity helps us particularly here. The notion of negative or positive sentiment comes from the fact that it is more negative/positive than the usual words used by people in ordinary contexts in daily life. There is no such thing as the universal notion of negativity or positivity. So, if we want to determine the negativity of a word, we should find how different it is (how much more negative) than the average word used in English. In order to accomplish that, we had to find the nature of our data and find what the properties of an average word are.

Table 1. Example of Scores from CDF

SentiWordNet Score	Objective prob. (%)	positive prob. (%)	negative prob. (%)
0	0	0	0
0.025	0.01	19.36	11.07
0.05	0.14	82.30	71.09
0.1	0.28	85.53	84.66
0.25	1.26	91.44	89.62
0.75	17.58	99.85	99.67
0.9	24.78	99.98	99.86
0.95	58.02	99.98	99.98
0.975	94.58	99.99	99.99
1	100	100	100

We made use of the Cumulative Distribution Function (CDF) Graphs in order to learn the aforementioned nature in our data. CDF of all words gave us the relative position of a word, which, in turn, gives us an insight of how much that word differs from the average word in dictionary. We extracted the score of every single word on SentiWordNet and using the word's score, we calculated the probability of being positive/negative as the probability of seeing a word with such a positivity/negativity score in a text by finding the underlying area of the curve in CDF. Table 1 shows a sample of outputs from the CDF graphs for each category. For instance, the word "speculation" has SentiWordNet scores of 0.675, 0.125, 0.250 for objectivity, positivity and negativity. Using the CDF based normalization, the scores for objective, positive and negative probabilities become approximately 0.15, 0.86, 0.90 respectively.

4 Experimental Results

We perform the sentimental measurements at the sentence-level, and then combine them to make the corpus-level decision of being factual or opinionated.

Table 2. Summary Statistics of Positive and Negative Scores at the Corpus Levels

	Editorials	NYT	Reuters
Mean of Positive Scores	0.32783	0.28363	0.22636
Median Positive Scores	0.32787	0.25194	0.23227
SD of Positive Scores	0.05323	0.09969	0.10252
IQR of Positive Scores	0.06816	0.12195	0.12154
Mean of Negative Scores	0.26938	0.20059	0.17295
Median of Negative Scores	0.26262	0.20885	0.17648
SD of Negative Scores	0.04395	0.08377	0.09523
IQR of Negative Scores	0.04807	0.10634	0.12238

To compute the degree of being factual/opinionated at the corpus levels, we calculated summary statistics of positive and negative scores for all the datasets. Table 2 shows the mean, median, standard deviation, and interquartile range (IQR) of sentences we get from the entire corpus. Mean-standard deviation and median-IQR values are useful pairs to observe the statistical characteristics. By the way, each sentence is taken into account independently without emphasizing which document they belong to.

If we take a look at the positive scores, we will see that the median score is gradually decreasing from 32.78 percent in editorials to 25.19 percent in NYT and finally 23.22 percent in Reuters, as expected. The same thing applies to negativity too. The probability of an average article to be negative is decreasing from 26.26 percent in editorials to 20.88 percent in NYT and finally to 17.64 percent in Reuters articles. There is almost 5 percent decrease in scores from editorial to NYT.

To test the significance of corpus-level positive and negative score differences, we applied analysis of variance on these different corpora. The analysis of variance (refer to Table 3) generates the F-score of negativity as 2370.54 and the F-score of positivity as 8765.38, and both of them lead to a p-value of $2x10^{-16}$. So, we can reject the null hypothesis, stating that editorials, NYT articles and Reuters are equal in objectivity.

Table 3. Results of Analysis of Variance

	F	p-value
Positive scores	8765.38	$2x10^{-16}$
Negative Scores	2370.54	$2x10^{-16}$

Furthermore, we applied Scheffé test to see how those three sources can be compared pairwise. By looking at Table 4, we can safely say that all three sources are fairly different from each other with p-scores lower than 0.05.

As a comparison base, we selected the unsupervised semantic orientation method by (Singh et al. 2013) as according to a recent survey (Dashtipour et al. 2016) it's the best state-of-the-art lexicon-based method in terms of its accuracy.

We performed first the analysis of variance on the semantic orientation scores of editorials, regular New York Times articles, and Reuters articles and then applied the Scheffé test to decide on the pairwise differences on being opinionated. The summary

Table 4. Results of Scheffé Tests

Comparison	F	p-value
Negative(Editorials vs. NYT)	7.99	$3.38x10^{-4}$
Negative(Editorials vs. Reuters)	15.65	$1.59x10^{-7}$
Negative(NYT vs. Reuters)	2355.56	0
Positive(Editorials vs. NYT)	2.82	$5.98x10^{-2}$
Positive(Editorials vs. Reuters)	14.85	$3.55x10^{-7}$
Positive(NYT vs. Reuters)	8751.78	0

statistics of the semantic orientation scores can be found in Table 5, and the analysis of variance and Scheffé test results can be observed in Table 6. The analysis of variance results show that the unsupervised semantic orientation method can make a separation among groups that are editorials, regular New York Times articles and Reuters articles in a statistically significant way. As for Scheffé test's results, the method fails to separate editorials from regular New York Times articles in a statistically significant way. For this pairwise assessment, our method is better as it does make the separation in a statistically significant way.

Table 5. Summary Statistics of Semantic Orientation Scores at the Corpus Levels

	Editorials	NYT	Reuters
Mean of SO Scores	11.06667	11.98192	3.012828
Median of SO Scores	11	6	1
SD of SO Scores	8.270985	17.17534	5.08865
IQR of SO Scores	9.5	17	4

Table 6. Test Results of Semantic Orientation Scores

		F	p-value
Analysis of Variance	SO scores	60088	$2x10^{-16}$
Scheffé Tests	SO(Editorials vs. NYT)		0.784695
	SO(Editorials vs. Reuters)		0
	SO(NYT vs. Reuters)		0

Although the aim of this paper is to find the factual/opinionated ratio of news sources, we add an additional experiment to see how it will perform under different circumstances, namely the binary document polarity task using the standard product review dataset (Blitzer, Dredze, and Pereira 2007). We had an overall accuracy of 52 percent due to 96 percent in positive reviews and 10 percent in negative ones. Our implementation of the unsupervised semantic orientation method, on the other hand, produced an overall accuracy score of 62 percent attributed to 73 percent in positive reviews and 53 percent in the negatives. It seems sufficient to compare against this work as it is the best lexicon-based method for this task and in general the second best after SVM with respect to the accuracy (Dashtipour et al. 2016).

5 Conclusion

We have presented a technique to evaluate news sources' degree of being factual or opinionated in an effort to measure their credibility. The technique assigns positive or negative sentiment scores to words using their cumulative distributions in a sentiment lexicon (SentiWordNet).

As we consider the cumulative distribution of the words in the lexicon, the scores we get are already normalized. Comparisons are done at the news source (corpus) level and each sentence is taken into account independently without emphasizing which document it belongs to. The technique's unsupervised statistical nature allows it to be applied along with any kind of sentiment analysis technique whether it is a lexicon-based model or a machine learning based model.

6 Future Work

Although the results are convincing, any future attempt at this subject should consider the context in which words are written. It is expected to perform much better since the objectivity and sentiment of a word can change drastically depending on the context. Furthermore, finding the context and the subject of the article (finance, sports, politics etc.) is important for any kind of analysis because it is acceptable to assume that in different areas of news, emotions are expressed in different ways. So, in future, topic modeling should be added to this method to calculate the shifts in positivity/negativity and improve the results.

In our project, we focused only one of the factors that leads to the credibility of news sources. However, in order to get the bigger picture it is necessary to investigate other factors, too. In this regard, we are also planning to add a second factor, "biased or unbiased", to our analysis of trustworthiness to cover the credibility of news sources more.

7 Acknowledgments

This paper is based on work supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under contract number 114E784.

References

[Bethard et al.(2004)] Bethard S, Yu H, Thornton A, Hatzivassiloglou V, and Jurafsky D (2004) Automatic Extraction of Opinion Propositions and their Holders. In: Proceedings of the spring symposium on exploring attitude and affect in text. AAAI, pp 22-24

[Blitzer et al.(2007)] John Blitzer and Mark Dredze and Fernando Pereira (2007). Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In In ACL, 187-205.

[Dashtipour et al.(2016)] Dashtipour, Kia and Poria, Soujanya and Hussain, Amir and Cambria, Erik and Hawalah, Ahmad Y. A. and Gelbukh, Alexander and Zhou, Qiang (2016). Multi- lingual sentiment analysis: State of the art and independent comparison of techniques. Cognitive Computation 8:757-771. 9415[PII] 27563360[pmid] Cognit Comput.

[Esuli and Sebastiani(2006)] Esuli A, and Sebastiani F (2006) SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation. LREC'06, pp 417-422

[Fahrni and Klenner(2008)] Fahrni A, Klenner M (2008) Old wine or warm beer: target-specific sentiment analysis of adjectives. In: Proceedings of the symposium on affective language in human and machine. AISB, pp 60-63

[Feldman (2013)] Feldman R (2013) Techniques and Applications for Sentiment Analysis. Commun. ACM 56(4): 82-89

[Fenby(1986)] Fenby, J.(1986). The international news services. Schocken Books.

[Gaziano and McGrath (1986)] Gaziano C, McGrath K (1986) Measuring the Concept of Credibility. Journalism Quarterly 63: 451-462

[Hogenboom et al.(2015)] Hogenboom A, Frasincar F, de Jong F, and Kaymak U (2015) Using Rhetorical Structure in Sentiment Analysis. Commun. ACM 58(7): 69-77

[Liu(2012)] Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 5(1): 1-167

[Miller(1995)] Miller GA (1995) WordNet: A Lexical Database for English. Commun. ACM 39-41

[Morinaga et al.(2002)] Morinaga S, Yamanishi K, Tateishi K, and Fukushima T (2002) Mining Product Reputations on the Web. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, pp 341-349

[Pang et al.(2002)] Pang B, Lee L, and Vaithyanathan S (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL-02, Association for Computational Linguistics, Stroudsburg, PA pp 79-86

[Pang and Lee(2008)] Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1-2): 1-135

[Pang and Lee(2012)] Pang, B and Lee, L and Vaithyanathan, S (2012).Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1-2):1-135.

[Rose et al.(2002)] Rose T, Stevenson M, and Whitehead M (2002) The Reuters Corpus Volume 1 - From Yesterday's News to Tomorrow's Language Resources, In Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC'2002, pp 827-832

[Sandhaus(2008)] Sandhaus E (2008) The New York Times Annotated Corpus Overview. https://catalog.ldc.upenn.edu/docs/LDC2008T1 9/new_york_times_annotated_corpus.pdf. Accessed 16 September 2014

[Singh et al.(2013)] V. K. Singh and R. Piryani and A. Uddin and P. Waila and Marisha (2013) Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches Knowledge and Smart Technology (KST), 2013 5th International Conference, pp 122-127

[Toutanova et al.(2003)] Toutanova K, Klei D, Manning C, and Singer Y (2003) Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL, pp 252-259

[Tsytsarau and Palpanas(2012)] Tsytsarau M, and Palpanas T (2012) Survey on Mining Subjective Data on the Web. Data Min Knowl Discov 24(3): 478-514

[Yu and Hatzivassiloglou(2003)] Yu H, and Hatzivassiloglou V (2003) Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, pp 129-136