Gold-standard for Topic-specific Sentiment Analysis of Economic Texts

Pyry Takala, Pekka Malo, Ankur Sinha, Oskar Ahlgren
Department of Information and Service Economy, Aalto University School of Business
P.O. Box 21210, FI-00076 Aalto, Finland
pyry.takala@aalto.fi, pekka.malo@aalto.fi, ankur.sinha@aalto.fi, oskar.ahlgren@aalto.fi

Abstract

Public opinion, as measured by media sentiment, can be an important indicator in the financial and economic context. These are domains where traditional sentiment estimation techniques often struggle, and existing annotated sentiment text collections are of less use. Though considerable progress has been made in analyzing sentiments at sentence-level, performing topic-dependent sentiment analysis is still a relatively uncharted territory. The computation of topic-specific sentiments has commonly relied on naive aggregation methods without much consideration to the relevance of the sentences to the given topic. Clearly, the use of such methods leads to a substantial increase in noise-to-signal ratio. To foster development of methods for measuring topic-specific sentiments in documents, we have collected and annotated a corpus of financial news that have been sampled from Thomson Reuters newswire. In this paper, we describe the annotation process and evaluate the quality of the dataset using a number of inter-annotator agreement metrics. The annotations of 297 documents and over 9000 sentences can be used for research purposes when developing methods for detecting topic-wise sentiment in financial text.

Keywords: Financial news, Corpus, Sentiment analysis

1. Introduction

Interest towards monitoring public opinion has grown during the past few years. Increasing news availability, the boom in social media, and recent technological advances have contributed to increasing analysis opportunities of media sentiment. In the financial context, media sentiment has often been linked to prediction of abnormal returns, volatility, and trading volume. Various arguments have been presented in favor of using media analytics alongside financial metrics. Recent studies in behavioral finance suggest that qualitative information, such as linguistic style and tone choices, may have an impact on investor behavior (Tetlock et al., 2008) (Loughran and McDonald, 2011). However, the underlying technologies are still in nascent state and the use of simple methods, such as wordcounts, is still surprisingly popular.

Though there is no scarcity of models for detecting polarity at sentence-level (Malo et al., 2014), considerably less research has been done towards understanding how sentence level sentiments determine the aggregated sentiment for a given topic at the document-level. One of the concerns is the topic-drift within a document, which means that topics may evolve and change througout the document. As a result, aggregation of sentence-level sentiments to estimate topic-specific sentiment at document-level becomes difficult. Not all sentences are equally relevant for the topic. An intelligent aggregation approach would take the relevance of the sentences into account while determining their effect on the topic-specific sentiment.

Sophisticated sentiment analysis is often performed using models that rely on supervised learning. These techniques require high-quality annotated datasets during training and performance evaluation. However, in spite of recent introductions (Loughran and McDonald, 2011; Malo et al., 2013b; Malo et al., 2014), annotated collections and language resources are still scarce in the field of economics and finance, and others are reserved for proprietary use only (O'Hare et al., 2009). To the best of our knowledge, there does not exist any publicly available dataset that provides topic-specific sentiments at both sentence and document levels.

To fill this gap, we introduce a human-annotated dataset that can be used for training and evaluation of topic-dependent sentiment models in the financial and economic context. The dataset consists of 10 different topics with around 30 news stories under each topic. All news stories have been taken from Thomson Reuters newswire in 2010. The chosen topics are diverse and range from company specific news to major events with macroeconomic impacts. Each news story in the dataset is annotated with the following information: (i) the key concepts featured in the story and their document-level sentiments; and (ii) the concepts featured within each sentence along with their sentence-level sentiments. The dataset was annotated by three independent annotators, who were carefully screened to ensure that they had sufficient economic and financial knowledge. All sentiment annotations have been done from an investor's perspective. The corpus can be used to understand how sentence level sentiment can be used to obtain an aggregated sentiment regarding a given topic from one or more documents.

The rest of the paper is organized as follows. Section 2 presents related work and contributions of the paper. Section 3 discusses the obtained dataset and evaluates the annotations. Section 4 concludes.

2. Related work

2.1. Sentiment analysis in finance

In news analytics, sentiment could be defined as the explicitly stated feelings, or it can also include emotions that arise from facts featured in the text (Balahur et al., 2010). For example, consider the sentence *"We just explained why the market currently loves Amazon and hates Apple."*, which expresses a positive sentiment towards Amazon and a neg-

ative sentiment towards Apple. Sentiment can be defined in a number of ways depending on the purpose and domain. The notion of domain dependence is particularly important in finance, where the sentiments have more to do with expected favorable or unfavorable directions of events from an investor's perspective.

Compared to the commonly analyzed domains such as movie reviews, financial texts often include more complex sentiment structures. One reason is the need to take expectations into account. For example, the sentiment of a statement "The losses dropped by 75% compared to last year." could be seen as positive. Making a loss is generally considered to be negative, but in this case the company has improved its business. In addition to the role of expectations, viewpoints should also be considered; e.g, news about layoffs could be seen as negative by the public, but as positive by profit-seeking investors. Also, many words that are considered polarized in common language (e.g. "tax") may be neutral in the financial context (Loughran and McDonald, 2011).

A number of studies have already been performed to find ways to measure sentiments in the financial domain. Many of them have relied on the use of "bag of words" methods (Engelberg, 2008; Tetlock et al., 2008; Garcia, 2013). Recently, there has been an increasing interest towards the use of statistical and machine-learning techniques to obtain more accurate sentiment estimates; e.g, naïve Bayesian classifier (Li, 2009), Support Vector Machines (SVMs) (O'Hare et al., 2009), multiple-classifier voting systems (Das and Chen, 2007) or domain-dependent classifier models (Malo et al., 2013a; Malo et al., 2013b). However, most of the research has focused on developing strategies to predict sentence level sentiments without paying much attention to the topic.

One of the primary reasons has been the lack of language resources that would support training and evaluation of classifiers for topic-specific sentiment detection. Few of the studies that have examined topic-specific sentiments have commonly done it by summing polarized words together using measures such as negativity percent (Tetlock et al., 2008; Engelberg, 2008), sum of positive and negative (+1 for positive, -1 for negative) (Das, 2010), or term-weighting (Loughran and McDonald, 2011). However, the drawback of these approaches is that they do not take the degree of relevance of different sentences appearing in a text into account. Consider, for example, sentences "Samsung phones are selling in record volumes" and "Oranges are selling in record volumes". Both sentences have an identical sentiment structure, but should have different weights based on their relevance to an investor in Samsung.

2.2. Language resources for sentiment analysis

To the best of our knowledge, there are no publicly available resources for topic-dependent sentiment analysis in the financial and economic context. However, given the increasing interest towards sentiment analysis, a number of other language resources have already been developed. Below we briefly review some of the publicly available resources:

Financial Polarity Lexicon
The resource consists of finance-specific word lists for sentiment analysis. In addition to positive and negative words, the lexicon includes new categories, such as uncertain, litiguous, strong modal, and weak modal word lists. (Loughran and McDonald, 2011)

Financial Phrase Bank
The dataset provides a collection of ∼5000 phrases/sentences sampled from financial news texts and company press releases, which have been annotated as positive/negative/neutral by a group of 16 annotators with adequate business education background. The data set is available for download at http://alcomlab.com/resources/ (Malo et al., 2013b)

SenticNet
This semantic resource has been created to support concept-level sentiment analysis. Its common-sense knowledge base utilizes both AI and Semantic Web techniques in order to improve recognition, interpretation, and processing of natural language opinions (Cambria et al., 2012).

MPQA Opinion Corpus
The corpus contains annotated news articles from a variety of news sources. All articles have been manually annotated for opinions and other private states. In addition to the corpus, the authors have also published a subjectivity lexicon that can be used for general sentiment analysis. (Wiebe et al., 2005)

JRC corpus of general news quotes
The resource consists of a collection English language quotations that have been manually annotated for sentiment towards entities mentioned inside the quotation. (Balahur et al., 2010)

Movie Review Data v2.0
This resource consists of 1000 positive and 1000 negative processed movie reviews. (Pang and Lee, 2004)

Sanders Twitter Sentiment Corpus
Free dataset that contains over 5500 hand classified tweets. Each tweet has been classified with respect to one of four different topics. (Sanders, 2011).

Harvard General Inquirer
The General Inquirer is a lexicon that attaches syntactic, semantic, and pragmatic information to part-of-speech tagged words. Although it was originally developed in the 1960s, it is still maintained[1].

Affective Reasoner
Affective Reasoner is a collection of AI programs. These programs are able to interact with subjects using a variety of means including speech recognition, text-to-speech, real-time morphed schematic faces, and music (Elliott, 1997).

Affective Lexicon
In this resource 500 affective words were examined. Special attention was given to the isolation of words that referred to emotions (Ortony et al., 1987).

[1] http://www.wjh.harvard.edu/ inquirer/

SentiWordNet

SentiWordNet is an enhanced lexical resource for sentiment analysis and opinion mining. The corpus associates each synset of WordNet with three different sentiment scores: positivity, negativity, objectivity (Baccianella et al., 2010).

ICWSM 2010 JDPA Sentiment Corpus

The corpus consists of blog posts containing opinions about automobiles and digital cameras. These posts have been manually annotated for various entities and the entities are marked with the aggregated sentiments expressed towards them in the document (Kessler et al., 2010).

3. Dataset for topic detection

In this section we introduce our annotated dataset that is based on ThomsonReuters newswire from 2010.

3.1. Creation of the corpus

The first step was to choose 10 news topics with significant financial impact[2]. Given the topic statements, the newswire was filtered for documents of suitable sizes that matched with the selected topics. The final choice was made manually to ensure that the selected stories were relevant and unique. In total 297 stories, evenly distributed over all topics, were chosen.

3.2. Annotation

Three annotators with previous experience from similar tasks were chosen to perform the annotation. They were instructed to read and annotate the news stories as if they were investors in the company that was defined in the topic statement. In case that no company was mentioned, the annotators were asked to annotate the news story for general market sentiment. All tasks were carried out using a 7-point scale from very positive to very negative. The annotators were also allowed to use ambiguous if necessary.

Before identifying the sentiments, the first task of the annotators was to read the entire story and find a maximum of three concepts that played a central role in the news story. After finding the key concepts, the annotators were asked to grade their sentiments on the document level. Next, this was repeated for all the individual sentences in the document. The annotators were strongly instructed not to utilize prior knowledge they had regarding the given news topics. Each story was annotated in isolation from all other news stories.

The corpus was divided into two parts; a common set consisting of 42 documents and the main corpus consisting of 255 documents. The common set was processed by all three annotators and the main corpus was split into three parts, each annotated by two reviewers. All annotators were independent with no prior affiliation with the research project.

[2]The major events chosen as topics for the corpus were Apple's iPad, the Icelandic ash cloud, BP's oil disaster in the Gulf of Mexico, EuroZone crisis, U.S. foreclosures problematics, GM's IPO, U.S. Securities Exchange Commission, U.S. TARP program, Toyota's quality problems in the U.S. and the United-Continental merger.

3.3. Evaluation metrics

To measure the reliability of the annotation scheme and examine the degree of agreement between the annotators, we conducted an agreement study with all three annotators using the common set of 42 documents.

A wide range of commonly used metrics have been used to evaluate the corpus. Below is a brief description of the key statistics.

Kappa statistic

The Kappa statistic is used for assessing interrater agreement between raters on categorical data. Kappa is defined as

$$\kappa = \frac{p_a - p_e}{1 - p_e}, \tag{1}$$

where p_a denotes the relative observed agreement for the raters and p_e is the hypothetical probability of a chance agreement. There are two well-known variants of the kappa statistic: Cohen's kappa and Fleiss' kappa. The statistics are rather similar, but differ in one crucial aspect. Fleiss' kappa allows the metrics to be calculated for several raters (Fleiss, 1981), whereas Cohen's kappa limits the number of raters to two (Cohen, 1960). Consequently, the way p_a and p_e are calculated differs. Landis and Koch proposed guidelines for interpreting the agreement measures for categorical data, their suggestion can be found in Table 1 (Landis and Koch, 1977).

Kappa	Agreement
<0.00	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

Table 1: Guidelines for interpreting Kappa-statistic

The Intraclass Correlation

The Intraclass Correlation, ICC, evaluates the reliability of ratings by comparing the variances of different raters on the same subject to the total variation across all raters and all subjects. For a comprehensive discussion on Intraclass Correlation, please refer to (Shrout and Fleiss, 1979) or (Müller and Büttner, 1994). The ICC statistics is defined as:

$$ICC = \frac{\sigma_\alpha}{\sigma_\alpha + \sigma_e}, \tag{2}$$

where σ_α is the variance between raters and σ_e is the variance within raters due to noise. ICC can be calculated for consistency (systematic differences between raters are irrelevant) and absolute agreement (systematic differences are relevant). Both are reported to highlight possible tendencies towards negativity or positivity by individual raters.

Robinson's A and the Finn coefficient

Robinson's A and the Finn coefficient are other popular indexes of the interrater reliability of quantitative data. For Robinson's A, D denotes the disagreement between the raters and similarly D_{max} denotes the greatest possible value of D (Robinson, 1957).

$$A = 1 - \frac{D}{D_{max}}. \quad (3)$$

When there's a low variance between raters, i.e. there's a high level of agreement, the Finn coefficient is particularly useful (Finn, 1970). In these cases the index can be calculated as:

$$r = 1.0 - \frac{MS_e}{(K^2 - 1)/12}, \quad (4)$$

where MS_e is the mean square error obtained from a one-way ANOVA and K is the number of scale categories.

The average percentage agreement

The average percentage agreement measures the ratio when all annotators agree.

All metrics calculations were done using R version 2.14.1[3] together with packages "irr" version 0.84 and "psy" version 1.1.

3.4. Evaluation of the annotation

Tables 2 and 3 summarize the sentiment distributions at the document and sentence levels, respectively. The difficulties with sentiment analysis are clearly highlighted when considering the label distribution coming from the different annotators. In particular these differences can be observed when examining the results on a 7-point scale. The annotators are generally found to agree on the direction, but they often have different opinions regarding the strengths.

7-point scale	Annot. 1	Annot. 2	Annot. 3
Very positive	14	0	0
Positive	14	3	8
Slightly positive	0	15	19
Neutral	0	17	1
Slightly negative	4	9	12
Negative	11	14	13
Very negative	15	0	5
3-point scale	**Annot. 1**	**Annot. 2**	**Annot. 3**
Positive	28	18	27
Neutral	0	17	1
Negative	30	23	30

Table 2: 7-point and 3-point scale sentiment distributions on the document level for the common set

Consider, for instance, the following snippet from a news story:

[3]http://cran.r-project.org/

7-point scale	Annot. 1	Annot. 2	Annot. 3
Very positive	14	43	2
Positive	39	38	10
Slightly positive	49	17	76
Neutral	10	8	25
Slightly negative	60	33	127
Negative	65	62	40
Very negative	43	79	0
3-point scale	**Annot. 1**	**Annot. 2**	**Annot. 3**
Positive	102	98	88
Neutral	10	8	25
Negative	168	174	167

Table 3: 7-point and 3-point scale sentiment distributions on the sentence level for the common set

> *NEW YORK, April 9 (Reuters) - A merger between United Airlines and US Airways would leave Continental Airlines in a tough spot, facing what some analysts called bleak prospects of flying solo among far-larger rivals.*

The corresponding annotator judgments for the snippet are given in Table 4. In this case, all annotators agree that the sentence is negative for Continental Airlines, but their opinions differ on the strength. In general, we find that the annotators agree more often on negative than positive news. In the case, the annotators had disagreements about the concepts themselves, all three concepts were discarded[4].

Sentence level	Concepts (sentiment)
Annotator 1	Continental Airlines (-2)
Annotator 2	Continental Airlines (-2)
Annotator 3	Continental (-1)

Table 4: Sentence level sentiment annotation for Continental Airlines

Given the strength related disagreements in the 7-point scale, we have chosen to simplify the annotations to a 3-point scale while presenting the statistics in the remaining parts of the paper. Table 5 shows multirater agreement statistics for 3-point annotation scale. For comparison, the results are also reported for a 2-point scale (positive, negative). If one of the annotators used neutral, then the entire concept was removed. Overall, the statistics suggest reasonably good agreement between the annotators. Both consistency and agreement versions of the ICC ratio are reported along with a number of other reliability measures.

The pairwise agreements and Cohen's kappa measures for the 3-point scale are given in Table 6 and 7. The results suggest a fair agreement especially on the sentence level

[4]If the concepts had the same meaning, but differed slightly in terms of the written forms chosen by individual annotators, the concepts were considered to match. For example "Continental" and "Continental Airlines" clearly link to the same concept and can thus safely be mapped without risk of corrupting of the corpus. On the other hand, it might be safe to map "the German carrier" to "Lufthansa", but in this case the entire sentence structure must be evaluated. The mappings of concepts were manually validated.

Document level	2-Point scale	3-Point scale
Fleiss' kappa	0.834	0.549
ICC Consistency	0.839	0.737
ICC Agreement	0.838	0.740
Robinson's A	0.893	0.825
Finn-Coefficient	0.851	0.641
Aver. Percent agreem.	0.919	0.736
Sentence level	**2-Point scale**	**3-Point scale**
Fleiss' kappa	0.901	0.721
ICC Consistency	0.901	0.842
ICC Agreement	0.901	0.842
Robinson's A	0.934	0.895
Finn-Coefficient	0.904	0.791
Aver. Percent agreem.	0.955	0.722

Table 5: Multirater agreements for the common set

Agreement	Annot. 1	Annot. 2	Annot. 3
Annot. 1	-	0.621	0.897
Annot. 2	0.621	-	0.690
Annot. 3	0.897	0.690	-
Cohen's kappa	**Annot. 1**	**Annot. 2**	**Annot. 3**
Annot. 1	-	0.412	0.796
Annot. 2	0.412	-	0.519
Annot. 3	0.796	0.519	-

Table 6: Pairwise agreement and Cohen's kappa and for the 3-point scale distribution on the document level for the common set

between the annotators as expected based on the multirater statistics. This is also to be expected since a single sentence seldom gives much reasons for huge disagreements. However, we also observe a few differences. Whereas the annotation patterns for annotators 1 and 3 are very similar as measured by Cohen's kappa, which is on the border of almost perfect agreement on the kappa scale, there is a relatively low correlation between these two annotators and annotator 2. The low correlation is mainly due to the frequent use of neutral label by annotator 2. This is, however, not too surprising, since drawing the line between slightly positive news and neutral news is often quite challenging in financial context. The differences between the percentual agreement and Cohen's kappa-values may seem large, and reason lies in the way Cohen's kappa is calculated. Since there is only three possible choices the likelihood of a chance agreement is high and thus any aberration will therefore be severely punished.

4. Conclusions

In this paper we have introduced a human-annotated dataset with topic-specific sentiments at sentence and document level. The corpus would be useful for the development of topic-based sentiment models that rely on high-quality training and evaluation data. The dataset has been annotated by three independent annotators with a reasonable degree of agreement. The resulting dataset will be published for academic and research purposes only. Due to licence agreement with Thomson Reuters, please contact the authors for access to the database.

Agreement	Annot. 1	Annot. 2	Annot. 3
Annot. 1	-	0.879	0.857
Annot. 2	0.879	-	0.836
Annot. 3	0.857	0.836	-
Cohen's kappa	**Annot. 1**	**Annot. 2**	**Annot. 3**
Annot. 1	-	0.756	0.728
Annot. 2	0.756	-	0.682
Annot. 3	0.728	0.682	-

Table 7: Percentual agreement and Cohen's kappa and for the 3-point scale distribution on the sentence level for the common set

5. References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment Analysis in the News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Cambria, E., Havasi, C., and Hussain, A. (2012). SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *AAAI FLAIRS*, pages 202–207.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Das, S. and Chen, M. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9):1375–1388.

Das, S. (2010). News Analytics: Framework, Techniques and Metrics. *SCU Leavey School of Business Research Paper No. 11-08*.

Elliott, C. (1997). I picked up Catapia and other stories: A multimodal approach to expressivity for 'emotionally intelligent' agents. In *Proceedings of the First International Conference on Autonomous Agents*, pages 451–457.

Engelberg, J. (2008). Costly Information Processing: Evidence from Earnings Announcements. SSRN Working Paper: http://ssrn.com/abstract=1107998.

Finn, R. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 30:71–76.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York.

Garcia, D. (2013). Sentiment during Recessions. *Journal of Finance*, 68:1267–1300.

Kessler, J., Eckert, M., Clark, L., and Nicolov, N. (2010). The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*. Accessed: 2014-03-03.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Li, F. (2009). The Information Content of the Forward-looking Statements in Corporate Filings - a Naïve Bayesian Machine Learning Approach. Working paper.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual Analysis, Dictionaries and 10-Ks. *Journal of Finance*, 66(1):35–66.

Malo, P., Siitari, P., and Sinha, A. (2013a). Automated Query Learning with Wikipedia and Genetic Programming. *Artificial Intelligence*, 194:86–110.

Malo, P., Sinha, A., Takala, P., Ahlgren, O., and Lappalainen, I. (2013b). Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis. In *Proceedings of IEEE International Conference of Data Mining workshops (ICDM SENTIRE)*. IEEE Press.

Malo, P., Sinha, A., Takala, P., Korhonen, P., and Wallenius, J. (2014). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the American Society for Information Science and Technology*, 65(4):782–796.

Müller, R. and Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, 3:2465–2476.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., and Smeaton, A. (2009). Topic-Dependent Sentiment Analysis of Financial Blogs. In *Proceedings of TSA '09*, pages 9–16.

Ortony, A., Clore, G. L., and Foss, M. (1987). The referential structure of the affective lexicon. volume 11, pages 341–364.

Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of ACL 2004*.

Robinson, W. (1957). The statistical measurement of agreement. *American Sociological Review*, 22:17–25.

Sanders, N. (2011). Twitter Sentiment Corpus. http://www.sananalytics.com/lab/twitter-sentiment. Accessed: 2014-03-03.

Shrout, P. and Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86:420–428.

Tetlock, P., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63:1437–1467.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.