Corporate News Classification and Valence Prediction: A Supervised Approach

Syed Aqueel Haider
Dept. of Computer Science & Engineering

MIT, Manipal University
KA-576104, India.
Aqueel.h.rizvi@gmail.com

Rishabh Mehrotra
Computer Science & Information Systems Group
BITS,Pilani
Rajasthan,India.
erishabh@gmail.com

Abstract

News articles have always been a prominent force in the formation of a company's financial image in the minds of the general public, especially the investors. Given the large amount of news being generated these days through various websites, it is possible to mine the general sentiment of a particular company being portrayed by media agencies over a period of time, which can be utilized to gauge the long term impact on the investment potential of the company. However, given such a vast amount of news data, we need to first separate corporate news from other kinds namely, sports, entertainment, science & technology, etc. We propose a system which takes news as, checks whether it is of corporate nature, and then identifies the polarity of the sentiment expressed in the news. The system is also capable of distinguishing the company/organization which is the subject of the news from other organizations which find mention, and this is used to pair the sentiment polarity with the identified company.

Introduction

With the rapid advancements in the field of information technology, the amount of information available has increased tremendously. News articles constitute the largest available portion of factual information about events happening in the world. Corporate news constitutes a major chunk of these news articles.

Sentiment Mining applied to the corporate domain would help in various ways like Automatic Recommendation Systems, to help organizations evaluate their market strategies help them frame their advertisement campaigns. Our system tries to address these issues by automating the entire process of news collection, organization/product detection and sentiment mining.

This paper is divided into two main parts. The first part describes a way of identifying corporate news from a collection of news articles and then pairing the news with the organization/company which is being talked about in the article. The second part of our paper works on the output of the first part (corporate news) and detects the valence of the identified corporate news articles. It calculates an overall score and identifies valence a s positive, negative or neutral based on this score. The system is immune to addition/mergers of companies, with regards to their identification, as it does not use any name lists.

The model uses a machine learning approach to do this task. We extract a set of features from the news and use them to train a set of classifiers. The best model is then used to classify the test data. One advantage of our approach described below is that it only requires a very small amount of annotated training data. We trained the model on the NewsCorp dataset consisting of 860 annotated news articles. The system has shown promising

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 175–181,
24 June, 2011, Portland, Oregon, USA

results on test data with classification accuracy being 92.05% and a f-measure of 92.00. The final average valence detection accuracy measured was 79.93%.

Related Work

Much work has been done on text classification.(Barak, 2009; Sebastiani,2002) There have been earlier attempts (Research on Sports Game News Information Extraction, Yonggui YANG,et al) However, they had focused mainly on information extraction and not classification.

Earlier attempts on web news classification(Krishnlal et al, 2010) concentrated mainly on classification according to the domain of the news articles. Not much work has been done in the field of corporate news-company pairing. This paper tries to address a more general problem of detecting the main organization being talked about in the articles.

Sentiment analysis in computational linguistics has focused on examining what textual features contribute to affective content of text and automatically detecting these features to derive a sentiment metric for a word, sentence or whole text. Niederhoffer (1971) after classifying New York Times headlines into 19 categories evaluated how the markets react to good and bad news.

Davis et al (2006) investigate the effects of optimistic or pessimistic language used in financial press releases on future firm performance. Sumbaly et al(2009) used k gram models to detect sentiment in large news datasets. Devitt(2007) improves upon and Melville(2009) have done work on sentiment analysis of web blogs

PART I : News Classification

Steps involved in news classification

3.1 News Pre-processing

The preprocessor merges all the files into one but defines start/end delimiters for each file in the merged file, to enable bulk processing. The merged news file is acted upon by a log-linear part of speech tagger we obtained from the Stanford NLP webpage(Manning,2000).

3.2 Organization detection

We follow a two step approach to organization detection:

Step 1: We extract the NNP/NNPS[1] clusters in the POS-tagged file using reguar expressions. For example, the pos-tagged version of "General Electric Co", is " General_NNP Electric_NNP Co_NNP" which is detected as a likely candidate for an organization.

Step 2: We use a Named Entity Recognizer[2] to obtain organization names. They are sorted in order of their frequencies and top three organizations are stored for later use. This ensures that even if some names have crept in as organizations due to misclassification by NER tagger, they end up at the bottom of the list and are discarded.

Multiple Organization Focus: Let f1,f2 be the frequencies of top 2 organizations. Now if f2>f1/2 then the news article is paired with organizations corresponding to both f1 and f2.

Baseline: Using just the frequency of top 3 organizations as features, we get an accuracy of 48.89% which is very low. Therefore, we add additional features which are described below.

3.3 Keyword Detection

The system matches each news article for occurrence of a set of keywords like "company", "share", "asset", etc. which have been derived from statistical observation of corporate news. We have used POS tags to differentiate between the contexts in which the keywords have been used. For example, "share" (verb) is not a keyword but "share" (noun) is a keyword. We calculate the net keyword occurrence frequency as $N(key)=\sum_{t=0} n(k_t)$ where N(key) is the total keyword frequency and $n(k_t)$ is the frequency of each keyword.

3.4 Headline Preprocessing

We process the headline and detect likely candidates for organization names and then cross check with the top 3 organization names detected in the step 2.2. We introduce a new feature h_value described as follows:

[1] Please refer Appendix A for details of the POS Tags.

$$h\_value = \begin{cases} \text{no. of matches in headline; if N(key) } > 5 \\ 0 \text{ , otherwise} \end{cases}$$

3.5 Detection of Products

The system detects likely candidates for products using three empirical rules:

- 1. _NNP followed by _POS followed by _NNP cluster. Ex: Google's ***Wave***
- 2.The followed by _NNP cluster. Example: The new ***POWER7*** processors from IBM
- 3._PRP$ followed by _NNP cluster. Example: Apple announced that its ***iPhone 3G*** will not be launched in India.

3.6 Executives Detection

We follow a similar POS based approach to detect executives, and store their frequency.

3.7 Feature Generation

We use a total of 9 features to train the SVM classifier. They are described below:
1-3: frequency of top 3 organizations
4: frequency of Executives in the news article
5-7: frequencies of top 3 products discussed in the news.
8. The N(key) value defined above in section 3.3
9. h_value defined above in section 3.2.

4 Classification and training

We tested our method with several classifiers.

First we used Support Vector Machines using LibSVM[**]. The results obtained were satisfactory. However, we experimented with other models to see model variation can lead to some improvement.

We tried **Logistic Regression** which is a class for building and using a multinomial logistic regression model with a ridge estimator. We trained our model with ridge parameter 1.0E-8.

We compared our classification results with **Naives Bayes** classifier which uses estimator classes for making the model. Numeric estimator precision values are chosen based on analysis of the training data.

We also tested our dataset with **AdaBoost** (Adaptive Boosting) classifier. AdaBoost calls a weak classifier repeatedly in a series of rounds to correctly identify the weights of the parameters.

The detailed results of the classification algorithms are discussed in the Experiments and Results section.

PART II : Headline Sentiment tagging

We describe a lexical features based approach to detect the sentiment polarity in a news article.

5.1 Preprocessing

One of the features of the news headlines extracted from the Internet was that many had all words capitalized. The system detects the improperly capitalized words and de-capitalizes their common words. This task is accomplished by using the following rule on the output given initially by the POS Tagger in Part I of our framework.

Rule: Only the words with POS tags as NNP or NNPS retain their capitalization, all others are decapitalized. Headline processing helps the POS Tagger to tag the words correctly and hence the dependencies will now be correct.

5.2 Stemming

Words which might carry opinions may be present in inflected forms which requires stemming of the words before any rules can be applied on them. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center. We have used the Porter Stemmer(Porter 1980)for this purpose.

5.3 Noise Reduction

The news article contains many parts of speech which are irrelevant to sentiment detection in our case, for example, prepositions, conjunctions, etc.

We give a list of Penn Treebank tags which we eliminate:

CC , CD, DT, EX, IN, PRP, PRP$, TO . Please refer to the Appendix A for the meaning of each POS-tag.

5.4 Polarity Estimation

We used the SentiWordNet (Sebastiani,2006) in order to calculate the sentiment polarity(valence) of all the words in the headline and the body.

We use WordNet to find sentiment polarity value(**SPV**) of each word. In WordNet, nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). Synsets represent terms or concepts. For example, following is a synset from WordNet:

stadium, bowl, arena, sports stadium – (a large structure for open-air sports or entertainments)

The synsets are related to other synsets higher or lower in the hierarchy by different types of relationships e.g.

- Hyponym/Hypernym (Is-A relationships)
- Meronym/Holonym (Part-Of relationships)
- Nine noun and several verb Is-A hierarchies

Using WordNet's word hierarchy we boosted sentiment polarities of a word (synset in WordNet), depending on whether a noun/verb, having a particular sentiment polarity is a hyponym of the given synset. The candidate synsets for polarity detection were extracted using a bootstrapping approach starting with some positive and negative seed words.

Parent synset	Boosted Polarity
poor	negative
good	positive
rise	positive
down	negative
decrease	negative
growth	positive
loss	negative

Table 1: Examples of hypernyms boosting sentiment polarity

5.5 Overall Valence Classification

After valences for each word have been detected, we proceed to find out the overall valence of the news article. We follow 2 rules for this task:

1. Since each word can have several meanings, to calculate the SPV of a word, we assumed that the these values were the average of all its possible meanings.

2. The SPV of words occurring in the headline are given higher weightage, as compared to those in the body. After several experimental trials, we concluded that a weight ratio of 4:1 was optimal.(4 for words in headline).

   The second rule is a direct consequence of the fact that news writers always try to provide the overall sentiment of the news in the headline itself so as to ease the understanding of the reader.

Now the overall valence score(OVS) is calculated using the simple expression OVS=

where SP is the Sentiment polarity value of each word in the news article.

Final decision:

$OVS > +k$,	positive polarity
$OVS < -k$,	negative polarity
$-k \leq OVS \leq k$,	neutral polarity

We experimented with different values of k and found out that a value of k=3 was most suitable for our task. Also, we could have normalized k according to the length of the news article to account for larger number of polar words in lengthier articles. However, we avoid doing so, because the probability of occurrence of positive polar words is the same as that of negative polar words, hence, neutralizing the effect of each other.

Finally, the OVS value provides a metric for the strength of the valence of news article. Higher magnitudes of OVS correspond to more strongly expressed sentiments.

6 Experiments and Results

In this section we discuss the dataset used in our experiments, the evaluation settings and the classification results obtained with our model.

6.1 The NewsCorp Dataset

We obtained 860 news samples from different news sites including:
1.ABC news
2. Reuters
3. MSNBC
4. CBC News Online, etc.

Our research team read these 860 news articles and created files for each of the news articles which contained details whether the article is corporate or non-corporate and if it is corporate then other details like main Organization being talked about in the article, different products and/or executives related to the organization mentioned in the article. We used these metadata files to evaluate our results regarding Organization, product and executive detection.
This dataset is then used to train the model for classification and also for sentiment mining task.

Sample metadata file:

```
<article>
        <headline>Apple Sells Three Million iPads in 80
Days</headline>
        <organization>
                <OrgName>Apple</OrgName>
                <product>iPad</product>
                <product>iPhone</product>
                <executive>Steve Jobs</executive>
        </organization>
        <sentiment>positive</sentiment>
</article>
```

6.2 Evaluation Methodology

We evaluate our method via 10-fold cross-validation, where we have sub-sampled the training folds in order to (a) keep the computational burden as low as possible and (b) show that we can learn sensible parameterizations based upon relatively low requirements in terms of the preferences seen on previous users. We evaluate the system in stages so that the contribution of each stage in the overall result becomes clear. We tested 860 news samples for corporate news detection. There were 261 true negative, 39 false positive, 83 false negative and 477 true positive articles. Precision, Recall and F-score are computed as:

Recall= TP / (TP+FN) Precision= TP / (TP+FP)

F-Score=(2.Precision.Recall)/(Precision+Recall)

We evaluated our results in three different stages. We first used basic Organization detection using NER tagger output as our baseline. Next we incorporated headline processing and keyword frequency detection in the second stage. Finally the third stage included the Product and Executive detection feature for result evaluation.

6.3 Classification Results

In order to classify the news article as corporate and non-corporate we used 4 different classification algorithms and compared their results. The four algorithms are:
1. Support Vector Machines
2. Logistic Regression
3. Naives Bayes
4. AdaBoost

Algorithm	Precision	Recall	F-Val	ROC Area
Naives Bayes	88.3	88.4	88.3	0.94
Support Vector Machine	85.81	92.44	85.17	0.94
Logistic Regression	90.4	89.9	90.0	0.95
AdaBoost	92.0	92.1	92.0	0.937

Table 2 (Classification Results)

Support Vector Machine gave us a third stage F Value of 88.66% while Naives Bayes gave a F Value of 88.3%.
Logistic Regression showed an improvement factor of 1.7% over Naives Bayes by giving F Value of 90.0%.
AdaBoost technique gave us the best classification result of 92& as the F value.
The different Precision, Recall, ROC Area and F Measure of the four algorithms are tabulated in Table 2 and Fig.2.

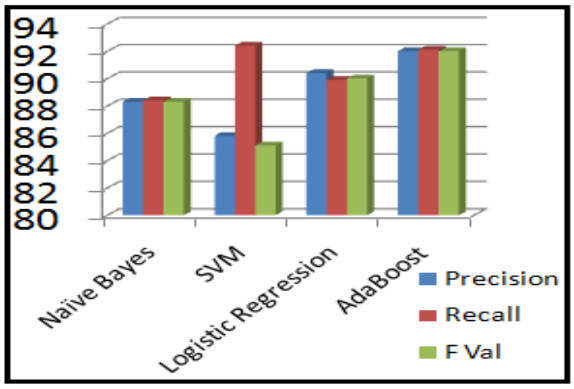

Fig. 1: Classification Results

6.4 Valence detection experimental results

The proposed system was tested with 608 articles since out of 860, 608 were identified to be of corporate type. The classification was 3 way, namely POS, NEG and NEUT (representing +ve, -ve and neutral respectively). The results are shown in Figure 1 in the form of a confusion matrix. Out of a total 608 financial news articles, 264 were tagged with positive sentiment, 162 with negative sentiment and 182 were found to be neutral.

		Predicted		
		POS	NEG	NEUT
Actual	POS	224	06	34
	NEG	04	148	10
	NEUT	50	18	114

Fig 2. Confusion Matrix for Valence Detection

However, our proposed approach yields an accuracy of 84.84, 91.35 and 62.35 for positive , negative and neutral news sentiments respectively . A possible reason for a low accuracy in case of neutral news articles could be because of the presence of some stray polar words in the body of the news, which might have added to a sum of more than 'k' in magnitude(as defined in Section 5.5), thereby leading to the development of an unwanted polarity.

Also, we observe a higher accuracy in predicting negative articles, the reason for which could not be identified. However, as proposed by a colleague, it could possibly be attributed to the fact that negative sentiment is more strongly expressed by Journalists in news articles, as compared to positive sentiment, which might have aided in better detection of words with negative polarity. Finally, we calculated the overall prediction accuracy by taking the average of accuracies for all three sentiments, which comes out to be 79.93%(Table 4).

	Precision	Recall	Accuracy
POS	80.58	84.85	84.84
NEG	86.05	91.46	91.35
NEUT	72.15	66.27	62.35

Table 3:Scores for Valence Detection

7 Conclusion and Future Work

A framework for valence identification and news classification has been proposed. News articles mined from the web by a crawler are fed to the system to filter the financial news from other kinds of news(sports, entertainment etc). Next, the organization which is the subject of this news is identified. Finally, we determine the sentiment polarity of the news by utilizing several lexical features and semantic relationships from WordNet.

We experiment with the system using our own manually tagged corpus of 860 news articles to fine tune various parameters like weights and threshold values. The resulting system performs well with identification of financial news as well as detection of valence in those articles. The system gives good result for positive and negative sentiments but satisfactory results for neutral sentiments. An overall accuracy of 79.93 % is obtained.

In the near future, we intend to apply anaphora resolution and use anaphoric distance to rank polar words according to relevance. This will help us to identify and give more weight to words which describe the sentiment of the author, from other "stray" words which are external references, not determining the overall sentiment of the news.

References

A New Text Mining Approach Based on HMM-SVM for Web News ClassificationLewis, D. D.: Reuters-21578 Document Corpus V1.0

Angela K. Davis, Jeremy M. Piger, and Lisa M. Sedor. 2006.Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. Technical report, Federal Reserve Bank of St Louis.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144{152. ACM Press, 1992.

Barak and Dagan.2009. Text Categorization from Category Name via Lexical Reference. Proceedings of NAACL HLT 2009.

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf

Devitt et al,(2007) Sentiment Polarity Identification in Financial News: A Cohesion-based Approach

George A. Miller (1995). WordNet: A Lexical Database for English.Communications of the ACM Vol. 38, No. 11: 39-41.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

Melville et al.(2009) Sentiment Analysis of Blogs by Combining. Lexical Knowledge with Text Classification.

Ozgur, A.: Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization. Master‘s Thesis (2004), Bogazici University, Turkey.

Porter, M.F. (1980) An Algorithm for Suffix Stripping, Program, 14(3): 130–137

Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34 no. 5 (2002)

Sentiment Mining in Large News Datasets. Roshan Sumbaly, Shakti Sinha, May 10, 2009.

UPAR7: A knowledge-based system for headline sentiment tagging. François-Régis Chaumartin Lattice/Talana – Université Paris 7

U. Hahn and M. Romacker Content Management in the SynDiKATe system — How t documents are automatically transformed to text knowledge bases. Data & Knowledge E ing, 35, 2000, pages 137-159.

Victor Niederhoffer. 1971. The analysis of world events and stock prices. Journal of Business, 44(2):193–219.

Appendix A. POS Tags

The POS tags used in Part I of the paper are described as follows:

NN = Noun
NNS = Plural Noun
NNP = Proper Noun
PRP = Personal Pronoun
PRP$ = Possessive Pronoun
JJ = Adjective
TO = *‘to’*
CD = Cardinal Number
DT = Determiner
CC = Coordinating conjunction
EX = Existential *there*
IN = Preposition or subordinating conjunction