International Journal of Computer Science and Applications

Vol. 9 No. 1, pp. 1 - 22, 2012

Classifying News Stories with a Constrained Learning Strategy to Estimate the Direction of a Market Index

Brett Drury
LIAAD-INESC, Rua de Ceuta, 118, 6
Porto, Portugal brett.drury@gmail.com
http://<www.liaad.up.pt>

Luis Torgo
Fac. Sciences, LIAAD-INESC, Rua de Ceuta, 118, 6
Porto, Portugal ltorgo@inescporto.pt
http://<www.liaad.up.pt>

J.J. Almeida
Depart. of Informatica,University of Minho, Braga, Portugal jj@di.uminho.pt
http://<www.uminho.pt>

News can contain information which may provide an indication of the future direction of a share or stock market index. The possibility of predicting future stock market prices has attracted an increasing numbers of industry practitioners and academic researchers to this area of investigation. Popular approaches have relied upon either: models constructed from manually selected training or manually constructed dictionaries. A potential flaw of manually selecting data is that the effectiveness of the trained model is dependent upon the ability of the human annotator. An alternative approach is to manually align news stories with trends in a specific market. A negative story is inferred if it co-occurs with a market losing value where as positive story is associated with a rise. This approach may have its flaws because news stories may co-occur with market movements by chance and consequently may inhibit the construction of a robust classifier with data gathered by this method. This paper presents a strategy which combines a: rule classifier, alignment strategy and self-training to induce a robust model for classifying news stories. The proposed method is compared with several competing methodologies and is evaluated with: estimated F-Measure and estimated trading returns. In addition the paper provides an evaluation of classifying a news story with its: headline, description or story text with: Language Models and Naive Bayes. The results demonstrate a clear advantage for the proposed methodology when evaluated by estimated F-Measure. The proposed strategy also produces the highest trading returns. In addition the paper clearly demonstrates that a news story's headline provides the greatest assistance for classification. The models induced from headlines gained the highest estimated F-Measure and trading returns for each strategy with the exception of the alignment method which performed uniformly poorly.

Keywords: sentiment; news; trading; semi-supervised learning; rules

1. Introduction

The inference of the direction of the price of a stock or share based upon the information contained in news stories has become an increasingly popular area of research for both the academic researcher and the commercial practitioner. News may contain timely information which can assist in the prediction of the prospects of economic actors [Mitra and Mitra(2011b)]. Information contained in news typically is concerned with the present or the future which is contrary to numeric data which typically describes the past [Kloptchenko et al.(2004)]. The potential of news information has generated a number of approaches which have had varying degrees of success. The published methods for predicting market returns based upon news information can be categorized as one of the following strategies: 1. manually created rules, 2. models learnt from manually selected data and 3. manually constructed dictionaries [Mittermayer and Knolmaye(2006)]. The methods have one central weakness, a reliance upon a human annotator. The effectiveness of the strategies may depend upon the annotator's ability to identify market moving stories.

Although news stories can contain timely information a delay in publication may devalue the information contained in the news story. Timely news delivered by newswire, for example Reuters or Bloomberg, is expensive. An alternative is to "scrape" news stories from the web, but only a fraction of news stories are transferred from news-wire to web sites. In addition there can be a significant delay for a news story to be transferred to financial web sites. The delay in publishing news may inhibit a strategy which attempts to align news stories with movements in the market to "bootstrap" training samples.

This paper describes a two step strategy which attempts to mitigate the problems of human annotation and alignment (with news stories published on the web) strategies which uses a combination of: 1. Manually constructed linguistic rules, 2. Unsupervised constructed dictionaries, 3. Alignment of stories with sharp market movements and 4. Self training on separate views of a news story to construct robust models to classify news stories. The strategy is tested with two classifiers: Naive Bayes and Language Models with two forms of evaluation: cross-validation and simple trading.

1.1. *Related Work*

The research literature has described a number of prototypes for predicting stock market's reaction to news. The first recorded approach was by the trader Victor Niederhoffer in the early 1970's. The stories from the day's newspapers were organised into 19 separate categories with a sliding polarity scale (positive to negative) [Taleb and Lane(2008)]. Trends were inferred from the aggregation of the polarity information. This manual approach would have been slow. The advent of machine readable news has allowed a number of systems to automatically classify news stories, eliminating the lag of Niederhoffer's approach. Thomas' [Thomas(2003)] Ph.D.

dissertation was arguably a descendent of Niederhoffer's strategy. Thomas created hand-crafted rules which assigned a news story to one of 39 categories, although it was noted by Mittermayer [Mittermayer and Knolmaye(2006)] that there was no published trading strategy. Wuthrich [Wuthrich et al.(1998)] attempted to classify stories which were published outside of market hours (over-nights). The classification strategy relied upon a manually constructed dictionary which contained 423 features. Mittermayer [Mittermayer and Knolmaye(2006)] reported that the dictionary was unpublished. A further system which relied upon a manually constructed dictionary was NewsCats [Mittermayer and Gerhard(2006)] which categorized press releases into three categories: (1. buy, 2. sell or 3. no-recommendation). An alternative to manually constructed rules and dictionaries is the alignment of news stories to market movements [Lavrenko et al.(2000)]. The alignment methodology infers that a news story is negative if it is published in the same time frame as a negative trend and a news story is positive if it co-occurs with a positive trend. There may be some flaws with this methodology: 1. news stories may be published out of time with market trends, 2. a news story may co-occur with a trend by chance, 3. a news story may contain contrary information to a market trend, for example a positive story co-occurring with a negative trend, 4. a market trend may be illusionary because the market can move without news information [Chan(2003)] . In addition Lavrenko [Lavrenko et al.(2000)] limited the methodology to stocks in single companies and selected stories where the company name was in the headline. Lavrenko's news selection methodology may be flawed because it ignored "spill over" [Mitra and Mitra(2011a)][a] effects from macro-economic and industry level news which can influence a stock's share price.

1.2. *Data Acquisition*

The proposed strategy required a large number of news stories for training and evaluation, consequently more than 300,000 stories were collected from a variety of news sites via Really Simple Syndication (RSS) feeds during the period from October 2008 until June 2010. The crawler ran at the same time each day. The following information from the RSS feed was stored in a database (RDBMS): headline, description, published date and story text. Meta data for the story text was provided by Open Calais [Wood(2010)]. Inter-day FTSE[b] stock price data for the same period was collected from Yahoo Finance.

1.3. *Layout of Article*

The article will address the following areas: 1.Overview of proposed strategy, 2. Linguistic rule induction (Domain Lexicon Construction, Grammar Induction, Evalu-

[a]Spill over refers to news stories which influences a company's share price although they don't explicitly reference the company
[b]http://www.ftse.com/

ation), 3. News story classification (competing strategies) and 4. Evaluation. The linguistic rule induction will demonstrate a methodology to generate JAPE[c] [Cunningham et al.(2002)] grammars which identify "interesting" phrases in news text. The news story classification section will describe a number of strategies to classify news stories. The evaluation section describes two evaluation criteria: F-Measure and Trading.

2. Overview of Proposed Strategy

The proposed strategy is a two-step process which attempts to identify "actionable" information in news stories which can be used in a trading system.

2.1. *Step 1- Bootstrapping Data*

The 1st phase of the algorithm is designed to return a pool of documents to use in the 2nd part of the algorithm which is a self-training phase. The 1st phase uses a combination of data which were selected by the alignment and rule strategies. This phase uses a voting strategy to ensure that a candidate news story must appear on a day where the market (FTSE) has either sharply: 1. decreased in value or 2. increased in value, and the candidate story conformed to a manually constructed linguistic rule.

The rational for the bootstrapping stages was to: 1. eliminate stories which appeared with a market movement by chance and 2. eliminate stories which could not be correlated with a significant market movement. The voting 'constraint' was intended to ensure that stories selected by both strategies had a higher probability of containing "actionable" information than stories which were individually selected by the aforementioned strategy. The 'bootstrapping algorithm' is described in Algorithm 1.

2.2. *Step 2- Self-Training*

The 2nd phase uses the pool of documents from step 1 to train base classifier(s) for a self-training process. Self-training uses a base classifier to select high confidence documents from a pool of unlabelled documents. These documents are then added to the previously labelled documents to induce a new learner. The new learner is then used in the next training cycle. This process continues until an explicit stopping condition is met or there are no further unlabelled documents [Abney(2008)].

The self-training is augmented by models induced from separate views of the news story. This is a process known as Co-Training [Blum and Mitchell(1998)]. The models were induced from the: Headline, Description (an RSS field) and the Story Text. The models in conjunction with the rule classifier would label the candidate stories. If a candidate story had the same label from each of the models and the

[c]JAPE is a system for manipulating annotations in the GATE platform http://goo.gl/4sVLe

Algorithm 1: 1st-Step Boot-Strap

Input: UL: A list of unlabelled stories
Input: Rise: Constant for minimal market increase
Input: Fall: Constant for minimal market decrease
** Linguistic Rule Classifier, a series of hand-crafted rules **;
$rc \leftarrow newRuleClassifier()$;
** Empty container to store stories and assigned label **;
$ld \leftarrow newHashMap()$;
** Main loop, add label if rule and alignment agree **;
forall $story \in UL$ **do**
  $marketmovement = MarketMovement(story.date)$;
  $label = rc.classify(story)$;
  if $(marketmovement \geq Rise \ and \ label ='' positive'')$ *or*
  $(marketmovement \leq Fall \ and \ label ='' negative'')$ **then**
    $ld.add(label, story)$;
return ld;

linguistic rule classifier it is added to the labelled set of documents to induce a new set of models. The self-training step is described in Algorithm 2.

3. Linguistic Rules Introduction

The proposed strategy is predicated upon manually constructed rules. The rules constructed for this work were reliant upon dictionaries constructed from the aforementioned corpus. This approach was preferred to the use of generally available sentiment dictionaries such as SentiwordNet [Esuli and Sebastiani(2006)] because these dictionaries may not contain domain specific language and in specific circumstances perform poorly [Balahur et al.(2010)]. In addition news stories may not contain sentimental language and therefore sentiment dictionaries may not assist in the detection of events which can effect the market. For example, the headline "American Airlines Enters Bankruptcy" may not contain sentimental language, but it can be interpreted as negative[d].

The rules were designed to capture and score phrases from textual information which contain either: 1. sentiment information or 2. "economic" event information. The rules were written in JAPE [Cunningham et al.(2002)] and model a phrase as a triple:1. Economic Actor (Company, Market ...), 2.Sentiment Adjective or Event Verb, 3. Economic Actor Property (Profits, Costs,...). An attempt is made to locate the target of the event or sentiment, and the rules employ strategies to score phrases when the "Named Entity"(Economic Actor) element is missing. The technique,

[d]The day the news story was published AMR share price dropped from a close of 1.62 to a close 0.26 on the next day of trading

Algorithm 2: 2nd-Step Self-Training Algorithm

Input: UL: A list of unlabelled stories
Input: LD: A list of labelled stories
Input: Const: A predefined constant for classifier confidence
** Induce base learners **;
$rc \leftarrow newRuleClassifier()$;
$hc \leftarrow TrainWithHeadlines(LD)$;
$dc \leftarrow TrainWithDescriptions(LD)$;
$sc \leftarrow TrainWithStoryText(LD)$;
$ul \leftarrow ()$;
$counter \leftarrow 0$;
forall $story \in UL$ **do**
    ** Classify each 'view' of the candidate news story **;
    $ruleC \leftarrow rc.classify(story.headline)$;
    $headC \leftarrow hc.classify(story.headline)$;
    $descC \leftarrow dc.classify(story.description)$;
    $textC \leftarrow sc.classify(story.text)$;
    ** Check classification confidence **;
    if $headC.conf < Const$ *or* $descC.conf < Const$ *or* $textC.conf < Const$ **then**
        $next$;
    ** Check classification agreement **;
    if $ruleC = headC \wedge headC = descC \wedge descC = textC$ **then**
        $LD \leftarrow LD.add(story, ruleC)$;
        $counter \leftarrow counter + 1$;
    else
        $ul \leftarrow ul.add(story)$;
** Termination, no further candidates **;
if $counter = 0$ **then**
    return LD;
else
    return $(SelfTrain(ul, LD, Const))$;

when tested against a small Gold Standard reports:1. a recall of 0.71 and 2. a precision of 0.94 for extracting sentiment phrases and a 0.84 recall and 0.83 precision for extracting event phrases. The remainder of this section will discuss in detail the construction and evaluation of the rules.

3.1. *Removal of Duplicate and Non-Financial News Stories*

The corpus contained in-excess of 300,000 news stories; although the stories were gathered from financial RSS feeds there were a large number of stories which were

non-financial. A number of stories were duplicated although they had different story url and publication dates. Duplicate stories were removed by the comparison of each story's RSS:headline and RSS:description fields with the existing stories's RSS:headline and RSS:description; if two stories or more had the same headline and description fields then all but one story was removed. A category for each news story was contained in the Open Calais Meta-Data [Wood(2010)]. If the news story was not categorized as financial or business news then it was removed from the training set. The remaining stories will be known as the "Training Stories".

The "Training Stories" text was split into sentences with the ANNIE Sentence Splitter [Cunningham et al.(2002)]. The following named entities were extracted from the meta-data: companies; organizations, market indexes and company employees. These types of entities for the purposes will known as financial named entities(FE). A sentence was removed from the training set if it did not contain one of the aforementioned entities. The process of removing sentences reduced the number of training sentences to approximately 500,000; this set of sentences will be known as "Training Sentences". 770 sentences were reserved and manually annotated with the following information: 1. Event phrase and direction, 2. Sentiment phrase and direction, 3. Sentiment / Event phrase target. This group of sentences will be known as the "gold standard".

3.2. *Domain Lexicon Construction and Analysis*

The identification of event phrases was predicated upon the discovery of event verbs; sentiment extraction relied upon the identification of opinionated words which the research literature indicates are normally adjectives [Indurkhya and Damerau(2010)]. The "Training Sentences" set was parsed with the ANNIE Part of Speech (POS) Tagger [Cunningham et al.(2002)] to assign part of speech information to each word in the "Training Sentences" set. The POS information was used to identify event and sentiment candidate words which will be described in the following subsections.

3.2.1. *Extraction of Event Verbs*

As previously stated verbs can be an indication of events [Levin(1993)]; consequently Verbs were extracted from the POS tagged "Training Sentences". The Verbs were extracted (base verbs) and sorted by frequency. The verbs which describe the actions of economic actors, for example the verb "rise" (Microsoft's shares will rise), will be known for the purpose of this paper as "event verbs". A subset of the "event verbs" which had a frequency of 2 or more were hand-selected from the base verbs, this set of verbs will be known as "base event verbs". The base event verbs were expanded with similar verbs from the Levin verb categories [Levin(1993)], for example the Verb "bounce" was part of the Roll Verbs category [Levin(1993)], consequently it was possible to expand "bounce" with the following words: "drift", "drop", "float",

"glide", "move", "roll", "slide", "swing". This expanded list will be known as "expanded event verbs". The "expanded event verbs list" was further expanded with semantic equivalents from Wordnet [Fellbaum(1998)] of each set member. This expanded list will be known as "final event verbs".

A domain expert assigned a type to the "final event verbs". A type was a label which was intended to the describe the function of the verb which provided illustrative feedback to the rule developer. The domain expert "scored" the "final event verbs". The "positive" verbs were assigned a symbolic score of "+1" and the "negative" verbs were given a nominal value of "-1" which reflected the positive and negative nature of the verbs. Examples of the verbs and their types are shown in Table 1. The types are for illustrative purposes only.

Verb Type	Examples
Obtained	gain(+1), add(+1), forge(+1), win(+1), attract(+1)
Lost	fire(-1), cut(-1), cancel(-1)
Direction	climb(+1), fall(-1), boost(+1), down(-1)
Behaviour	storm(+1), unravel(-1)
Influence	hurt(-1),hit(-1) push(+1), suffer(-1)

Table 1. Sample Verb Extraction Categorization

3.2.2. *Extraction of Sentiment Adjectives*

Sentiment information can be contained in adjectives [Benamara et al.(2007)], for example: 1. worst (negative), 2. good (positive). Adjectives were consequently extracted from the POS tagged training sentences. The adjectives were scored with information from Sentiwordnet. The adjectives were ranked by frequency and category. The adjectives which had a frequency of 2 or more were selected. The adjectives were checked by a domain expert and adjectives which did not correspond to their category were removed. The remaining adjectives will be known as the "seed adjective list".

The "seed adjective list" was expanded by extracting semantic equivalents from WordNet [Fellbaum(1998)]. This expanded list was further extended with algorithm 3, which is a simplified version of the algorithm described by Hatzivassiloglou and Columbia [Hatzivassiloglou and McKeown(1997)] which uses connectives to identify and predict new sentiment words and their orientation. Connectives were used in this instance to identify adjectives with the same sentiment label. For example: if the adjective "good" has a known polarity label (+) it is possible to propagate its label to other words. For example, the following sequence was extracted from our corpus: 1.good and cost-efficient, 2.cost-efficient and fair, 3.fair and transparent. In this sequence the label was propagated from "good" to:1. cost-effective, 2. fair,

3.transparent with the connective "and".

Algorithm 3: Description of Sentiment Propagation

Input: SL: A list of adjectives with sentiment labels
Input: UD: List of unlabelled sentences
Input: LC: List of Connectives
Output: SL: Expanded list of adjectives
** Calculate Word Trigrams in Sentence **;
$trigs \leftarrow calctrigrams(UD)$;
repeat
  $candidates \leftarrow \{\}$;
  for $(word, lab) \in SL.words$ **do**
    ** Align trigrams with known adjective and connector **;
    for $\{(w_1, w_2, w_3) \in trigs \mid w_1 = word\}$ **do**
      next if $w_2 \notin LC$;
      next if $w_3 \in SL$;
      $push(candidates, (w_3, lab))$;
  $SL \leftarrow SL \cup candidates$;
until *No more new candidates* ;

Each iteration of the algorithm produced new opinion words which were not in the input list of words; the new words were expanded with semantic equivalents from WordNet. A new input list which consisted of the newly expanded words was created and used as a seed list for the new iteration of the algorithm. This process was continued until no more new words were produced. The positive and negative words were assigned scores of "+1" and "-1".

3.2.3. *Extraction of Entity Features*

The initial set of training sentences was reduced by eliminating sentences which did not contain either: an event verb, a sentiment adjective. This new set of sentences will be known as the "reduced training set". The "reduced training set" was used to identify words which had a statistically significant relationship with the identified event verbs or sentiment adjectives. The sentences contained in the "reduced training set" had the following word types removed: stop words, proper nouns, named entities. The remaining words were extracted and labelled with one of the following categories: co-occurred with event verb, co-occurred with sentiment adjective, co-occurred with both sentiment adjective and event verb. The resulting words will be referred to as the "co-occurring word list".

A Pointwise Mutual Information (PMI) score was calculated for each member of the "co-occurring word list". The PMI score was computed for each member of the

"co-occurring word list" to assign an affinity with one of the following categories: 1. Co-occurrence with an event verb, 2. Co-occurrence with an sentiment adjective. The PMI equation is described in Equation 1. The *wrd* symbols represents the co-occurring word and *cl* represents a group of words which was either: the event verbs or the sentiment adjectives.

$$PMI(wrd, cl) = \log_2 \frac{Pr(wrd, cl)}{Pr(wrd)Pr(cl)}. \quad (1)$$

The words which had a PMI of 0 or less were removed because they were assumed to have co-occurred by chance. The collection of the remaining words will be referred to as "statistically significant word list". The members of the "statistically significant word list" were expanded with JSpell [Almeida and Pinto(1995)] to include all possible word forms. A sample of the members of the "statistically significant word list" can be found in Table 2 with arbitrary assigned types which are for illustrative purposes only.

Word Types	Examples
Success Measures	footfall, sales, profits, demand
Time Periods	Monday, Tuesday, January, month, year, period
Third Parties	investors, analysts, investors, economists, regulators, consumers
Miscellaneous	transactions, finance, bankruptcy

Table 2. Sample Features (Nouns)

3.2.4. *Extraction of Sentiment and Event modifiers*

The strength of a "sentiment word" may be modified by an adverb [Benamara et al.(2007)]. Sentiment modification maybe be one of the following: sentiment maximization (e.g. very good), sentiment minimization (e.g. fairly good) and negation (e.g. not good). The same procedure for identifying co-occurring nouns was used to identify sentiment modifying adverbs. The adverbs were hand-scored. The negation words with -1 because this would invert the score of a sentiment adjective in an extracted phrase. In the absence of a commonly accepted standard for scoring sentiment modifiers an arbitrary score of between 0.1 and 0.9 were applied to minimizers. The score would reduce the initial sentiment adjective score in an extracted phrase, for example a minimizer with a score of 0.9 would reduce the initial score

by a tenth. The maximizers were assigned a value between 1.1 and 2. This strategy sought to increase the score of sentiment adjectives in an extracted phrase. A sample of negation, maximizer and minimizer words are contained in Table 3.

Sentiment modifier types	Examples
Maximization	sharply, super, perfectly
Minimization	rickety, piffling, just
Negation	not, none, never

Table 3. Sentiment Modification (Adverbs)

3.2.5. *Lexicon Construction Summary*

The above process produced the following: 1. 2519 adjectives with a polarity label, 2. 393 verbs with a polarity features, 3. 2609 entity features, 4. 90 sentiment modifiers.

3.3. *Grammar Rule Induction for Event and Phrase Annotation*

This paper has thus far described the identification of words which have statistical relations with either an: Event Phrase or Sentiment Phrase. The motivation of this work was to annotate event or sentiment phrases, consequently it was necessary to construct a series of grammars (rules). The grammars were expressed in JAPE [Cunningham et al.(2002)], which can manipulate annotations in the GATE framework[e].

3.3.1. *Gate Annotations*

JAPE Grammars require annotations to manipulate. Words can be annotated in GATE with the GATE Gazetteer, the GATE Gazetteer holds a list of pairs: a word and it's annotation label. The Gazetteer was supplemented with the following: adverbs (sentiment modifiers), event verbs (verb); sentiment words (adjective) and statistically significant words (FE features) as new word lists. The Gazetteer already contained a company list which was expanded with FEs from the Open Calais [Wood(2010)] meta-data, which was collected in the data acquisition stage. The elements of event and sentiment phrases were now annotated in the story text and could be manipulated by JAPE.

3.3.2. *Phrase Extraction Patterns*

The JAPE grammars were based upon extraction triples. The JAPE rules were not order dependent, for example the event extraction pattern could be: Named

[e]http://gate.ac.uk/

Entity, Verb, Feature (Microsoft(NE) reported a drop(Verb) in profits(Feature)) or Named Entity, Feature, Verb (Microsoft's(NE) profits (Feature) dropped(Verb).). The grammars ensured that the longest possible phrase was returned.

(1) Third Party Financial Entities: There were two types of FEs: 1. The subject of the phrase or 2. A third party passing commenting. Named entities which have a linguistic cue that they were a third party, for example "Analysis by Lane Clark & Peacock, the actuarial consultants" [Pfeifer(2009)], were excluded from analysis.
(2) Partial Patterns and Backtracking: A number of event and sentiment phrases in the corpus did not contain a FE (company, market index, etc.). The title of the named entity would either be: implied, substituted with an informal name. Two strategies were followed to compensate for a missing FE: backtracking and partial patterns.

3.3.3. *Backtracking*

Frequently journalists replace company names in text which makes numerous references to the same named entity (company, market indexes, etc.) to ensure that the text is not repetitive. The backtracking strategy looked for an explicit reference to a named entity, which was not a third party passing comment, in the previous sentence. The annotated phrase was not expanded, but the inferred named entity was used as the event or sentiment target. In certain circumstances the partial phrase was expanded with partial pattern combination.

3.3.4. *Partial Patterns*

When the FE element was missing from the immediate sentence, the remaining elements (Verb and Feature or Adjective and Feature) were returned. The complete event and sentiment phrase was returned by combining two or more partial patterns in the same sentence. There were two combination rules:

Rule 1 - Partial patterns were joined when there was one separator token (space;newline,tab, etc) between partial patterns.
Rule 2 - Partial patterns were joined when they were separated by a "continuation" [Barker(2004)].
Note - The patterns must have the same "type", event phrases can't be joined with sentiment phrases.

Table 4. Combination Rules

Phrase	Classification
union hits out as lloyds bank< *EconomicActor* > axes< *Verb* > more jobs< *Object* >	negative
lukoil< *EconomicActor* > wins< *Verb* > west qurna-2 contract< *Object* >	positive
Stonefield Software< *EconomicActor* > Scores< *Verb* > Incredible< *adjective* > 42 Percent Sales< *Objective* > Growth	positive
Downturn and poor< *adjective* > weather wipe< *verb* > out C&C< *EconomicActor* > profits< *object* >	negative

Table 5. Examples and Classification

3.3.5. *Annotation of Complete Phrases*

Thus far there have been two types of patterns described: complete patterns and partial patterns. Complete patterns and combinations of partial patterns may capture whole event or sentiment phrases. On occasions these patterns may not be sufficient, consequently it was necessary to combine patterns, both complete and combined partial patterns. The combination rules were the same as for partial patterns as described in Table 4 .

3.3.6. *Estimating the Polarity of a Phrase*

The polarity of event and sentiment phrases were estimated separately. The event phrases were scored by assigning the polarity of the event verb to the whole phrase. The verbs were pre-assigned a value by a domain expert. A number of features acted as a negator, for example: "A rise in profits" would be positive whereas "A rise in costs" would be negative. The sentiment phrases were scored with the AVAC algorithm [Subrahmanian and Reforgiato(2008)]. The algorithm uses combinations of Adjectives, Verbs and Adverbs to estimate a score for the sentiment phrase. The sentiment examples in Table 5 show combinations of verbs and adjectives.

3.3.7. *Evaluation*

The initial evaluation was with a manually annotated "gold standard". The "gold standard" document collection was annotated by a single annotator. The "gold standard" evaluation was to determine: 1. precision and 2. recall of the rule based system with an expert's annotations. A trading evaluation was also conducted. This evaluation will be described later.

The evaluation of the phrases annotated by the grammars was by comparison with the previously mentioned the "gold standard" set of sentences. The evaluation consisted of the following tasks: 1. Correct identification of the sentiment or event

phrase, 2, Differentiation of an event from sentiment, 3. Correct identification of sentiment / event target and 4. Direction of sentiment or event. The rational of the evaluation criteria is described in Table 6.

Phrase extraction	Extracted phrase must convey the fundamental message of the annotated phrase
Event / sentiment detection	Must determine between a sentiment or event phrase. If annotation is uncertain then either will be accepted
Target of sentiment	Extracted sentiment target must exactly match the annotation
Sentiment / event direction	Extracted must match the direction of the annotation

Table 6. Evaluation Criteria

Evaluation Item	Recall	Precision
Sentiment phrase extraction and direction	0.71	0.94
Event phrase extraction and direction	0.84	0.83
Sentiment Target Extraction	0.74	0.74
Event target extraction	0.84	0.77

Table 7. Recall and Precision for Phrase Extraction

3.3.8. *Results*

The results of the evaluation are presented in Table 7. The assigned direction of the extracted sentiment phrase was correct for 77% of the extracted phrases, and the assigned direction of the extracted event phrase was correct for 74% of the extracted phrases.

4. Alignment of Market Data

The proposed strategy was predicated upon a second strategy which attempted to align news stories with large fluctuations in the market(FTSE). The rational behind this strategy was a large fluctuation indicated a significant economic event which would be published in the mass-media. A selection of events and market movements is presented in Table 4.

The stories were labelled either positive or negative if they were published on the same day as a significant market movement. For the purposes of this paper a single

Date	FTSE (+/-)	Reason
8th August 2011	-3.39%	Falls in US and Asian Markets
10th/12th Sept 2011	-2.73%	Terrorist Attacks
7th Sept 2008	-1.93%	Financial Crisis

Table 8. Large Fluctuations in the FTSE

day market movement is the difference between the opening and closing price on a single day. The proceeding and previous day are ignored. For example, a selection of headlines of stories published on 7th September 2008 (market fell by -1.93%) were: "Lehman shuts subprime arm and axes jobs" (http://goo.gl/WVjmC), "Mortgages to rise as crisis grips the markets" (http://goo.gl/GdTHn) and "US bank jobs go as loan crisis worsens" (http://goo.gl/HAuVU). These stories would have been labelled negative because the stories were published on a day were the market dropped. The labelling strategy used the 50 biggest drops and rises of the FTSE in the period 2008-2009. The news stories which were published on these days were used to induce a classifier. The alignment strategy was used in conjunction with the linguistic rules classifier in the first part of the proposed strategy to bootstrap the initial training data.

5. News Classification

The proposed algorithm was designed to classify news stories into pre-assigned polarity categories. The polarity categories represent the proposed effect of the news story on the market, consequently a positive news story should induce a rise in the market whereas a negative story should induce a drop in the market. The proposed algorithm was evaluated by: Estimated F-Measure and Trading Returns against several competing strategies. Two strategies (1. Proposed Algorithm and 2. Alignment Strategy) have been described in detail and the following sections describe the remaining news story classification strategies.

5.1. *Model Constructed from Rule Selected Data*

There are two rule based strategies in this article: 1. the linguistic rule classifier and 2. a model induced from rule selected data. The strategy which uses rule selected data to induce a model will be known as Rule Trained.

The rule trained data uses the linguistic rule classifier to label headlines of news story as either positive or negative. Abdication or classifications with a score of zero were ignored. This data was used to induce a classifier.

5.2. *Hybrid of Rules and Alignment*

This strategy is the first part of the proposed strategy,i.e the "bootstrapping step". The data is labelled as positive or negative if a news story is published on the same

day as a significant market movement and the headline conforms to a manual rule. The algorithm is described in Algorithm 1.

6. Evaluation

As stated earlier the evaluation methodology was two-fold: 1. an estimated F-Measure and 2. trading returns. The estimated F-Measure was estimated with Lingpipe's implementation of cross-validation with Language Models and Naive Bayes [Alias(2008)]. Two classifiers were used to ensure that any gain in F-Measure was not learner specific. The F-Measure was estimated with a 2 x 5 cross validation. The F-Measure was estimated for models generated from the story's: headline, description or story text. The results are presented in Tables 1 and 4.

The estimated F-Measure for both the Naive Bayes and Language Model classifiers indicates that the highest F-Measure was generated from headlines and the lowest scores were estimated from story text information. The alignment strategy was an exception, but the strategy performed uniformly poorly. There was an anomaly with the proposed strategy where the model induced by the description information for the Naive Bayes Classifier had a lower estimated F-Measure than the model induced from story text information, but the standard deviation was higher. The differences between the highest and lowest F-Measures are presented in Figure 6. The proposed method had a lower absolute and relative difference than the rule strategy, but a higher difference than the alignment and hybrid strategy, however the alignment and hybrid strategies had a significantly lower starting point.

The precision and recall estimated for the Language Model classifier is presented in Tables 2 and 3. The estimated figures demonstrate that proposed method gained both precision and recall through the self-training process. For example, the initial training set generated an estimated recall and precision of 0.66 from headline information. These estimations were increased through the training process to a precision of 0.86 and a recall of 0.83.

F-Measure:

Classifier	Headline	Story Text	Description
Alignment	0.57 ± 0.01	0.57 ± 0.01	0.57 ± 0.00
Hybrid	0.66 ± 0.04	0.57 ± 0.06	0.58 ± 0.04
Rule Trained	0.77 ± 0.01	0.60 ± 0.01	0.65 ± 0.01
Proposed	0.84 ± 0.01	0.71 ± 0.01	0.77 ± 0.01

Fig. 1. Estimated F-Measure for competing strategies (Language Models)

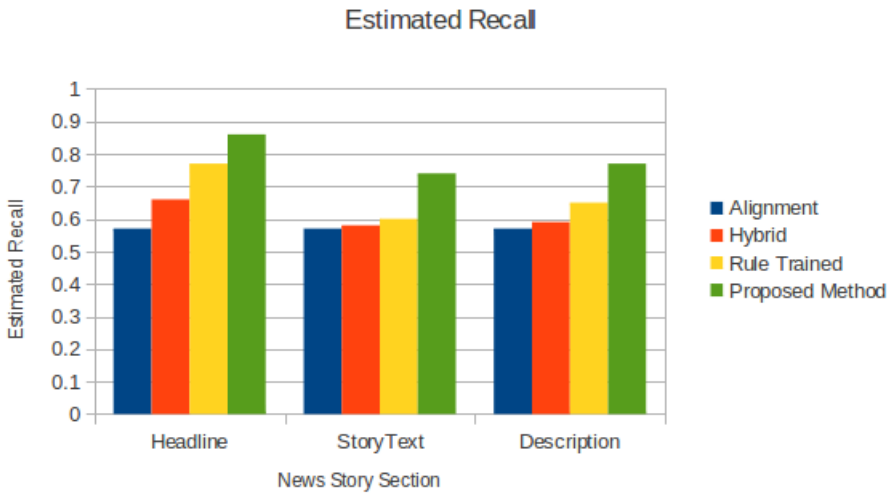

Fig. 2. Estimated Recall for Competing Methodologies (Language Models)

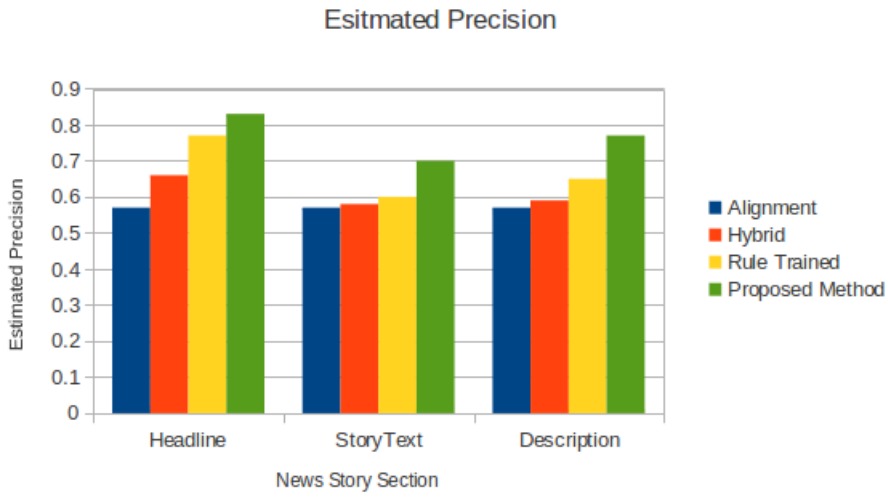

Fig. 3. Estimated Precision for Competing Methodologies (Language Models)

7. Trading Evaluation

This section will describe a secondary evaluation. The evaluation will estimate the percentage returns gained by trading on the FTSE-100 index based upon recommendations from classifiers trained by the competing strategies. It may be possible that the estimated F-Measure for the selected training data may not be an accurate

F-Measure:

Classifier	Headline	Story Text	Description
Alignment	0.57 ± 0.01	0.58 ± 0.01	0.57 ± 0.01
Hybrid	0.68 ± 0.05	0.58 ± 0.00	0.58 ± 0.02
Rule Trained	0.76 ± 0.01	0.60 ± 0.02	0.66 ± 0.00
Proposed	0.92 ± 0.00	0.77 ± 0.06	0.75 ± 0.13

Fig. 4. Estimated F-Measure for competing strategies (Naive Bayes)

Strategy	Classifier	Relative Difference	Absolute Difference
Proposed	Naive Bayes	18%	0.17
Proposed	Language Models	15%	0.13
Rule Trained	Naive Bayes	21%	0.16
Rule Trained	Language Models	22%	0.17
Hybrid	Naive Bayes	15%	0.10
Hybrid	Language Models	14%	0.09
Alignment	Naive Bayes	2%	0.01
Alignment	Language Models	0	0

Fig. 5. Difference between lowest and highest F-Measure

indicator of its effectiveness in estimating market movements. There is a caveat to this form of evaluation. The classifiers will have access to all the news stories published on a given day and consequently the evaluation will measure the effectiveness of the classifiers ability to classify a trading day as negative or positive by analysing news. A strategy which returned a positive return may not be guaranteed to gain a return in a trading environment. In a trading environment a trader will only have access to news published up to the time of trade where as in this evaluation the strategies will have access to the complete day's news when the market opens.

7.1. *Experimental Setup*

The data from 2008 and 2009 was reserved for training or for dictionary construction for the rule strategies. The evaluation data was news stories published in 2010. The evaluation determined that a day was either: negative or positive by counting the number of stories of each category on a given day. If there was at least a difference of 5 between the categories then the recommendation would be the dominant category. If there was no clear difference then the day would be determined to be neutral. The "trading strategy" would be: for a "positive day" to buy at the opening price and sell at the end of the day and 2. "short" on a negative day. Shorting implies selling at the opening price and buying at the closing price. On a neutral day no trading action would be undertaken.

Strategy	Voting	Headline	Description	Story Text
Alignment	-12.2%	-24.5%	-20.6%	-0.1%
Hybrid	-10.6%	-10.6%	-10.6%	-10.6%
Rule Trained	16.5%	16.8%	14.8%	-1.2%
S.T. Hybrid	-10.6%	33.8%	-6.5%	-12.3%
Rules (Event + Sentiment)	NA	-5.45%	-0.09%	3.11%
Rules (Event)	NA	-3.11%	-11.8%	-5.12%
Rules (Sentiment)	NA	14.83%	-4.87%	12.54%

Fig. 6. Trading Evaluation for All News Stories

Models were induced from: headline information, description text and story text. The evaluation task evaluated each classifier and in combination. The combination of classifiers, classify stories with a majority vote, therefore only two classifiers were required to agree. The rule methods were separated into three competing strategies: events, sentiment and sentiment and events. The rules classified headlines as a single phrase and summed extracted phrases to classify description and story text information.

There were two evaluation cycles: 1. count all classified stories and 2. count only classified stories with a high confidence (>90%). The competing methods were evaluated with a simple metric: the percentage "profit" made through trading activities. The results are presented in Tables 6 and 7. The majority of strategies made losses with the exception of: 1. Rule trained classifier, 2. the proposed method and 3. Sentiment rule classifier. The proposed method returned the highest "trading profits" when classifying stories with only headline information, but made losses when using either: story text or description information. The use of high confidence classifications increased returns. The rule trained and the sentiment rule classifier was the proposed method's nearest competitors because they made consistent positive returns. The rule trained classifier was the only classifier in the first evaluation cycle to make a positive return when classifying news stories by description information, but the classifier made negative returns in the second evaluation cycle where high confidence classifications were used for "trading". The sentiment rule classifier was the only classifier to make a trading profit from information in the story text.

7.2. *Evaluation of News Story Characteristics*

The results presented in Tables 4,1, 6 and7 demonstrate clearly that models induced from headline information construct are the most robust. The models induced from headlines consistently gained the highest F-Measure for each competing strategy as well highest returns or the lowest losses. The models induced from description text consistently performed worse than headline classifiers, but better than models induced from story text. This is reflected in trading returns and estimated F-Measure.

Strategy	Voting	Headline	Description	Story Text
Alignment	-12.4%	2.2%	-19.6%	-3.8%
Hybrid	-10.6%	-11.9%	-19.6%	-10.6%
Rule Trained	18.6%	30.0%	-10.6%	-5.9%
S.T. Hybrid	6%	47.2%	0.5%	-10.6%
Rules (Event + Sentiment)	NA	-5.45%	-0.09%	3.11%
Rules (Event)	NA	-3.11%	-11.8%	-5.12%
Rules (Sentiment)	NA	14.83%	-4.87%	12.54%

Fig. 7. Trading Evaluation for High Confidence Classifications

Models induced from story text were the weakest both in F-Measure and trading returns.

8. Conclusion and Future Work

A reoccurring problem for systems which attempt to classify news stories is the lack of training data. A further problem is the correct identification of stories that have market influence. Manual classification of stories or manual construction of dictionaries may be a long and laborious process which may yield insufficient training data or incomplete dictionaries. Alignment of news stories with market movements may assist in the selection of news stories for training data, however this method has its flaws. News stories may co-occur with market movements by chance or the news stories may be contrary to a market trend or movement.

This paper presents a proposed method for categorizing news stories into positive or negative categories. A rule classifier selects stories which have an event or sentiment phrase in its headline. These selected stories are then aligned with market movements. Stories which have the same label assigned by both strategies may limit the possibility of a story co-occurring by chance or contains contrary information to the trend. This method may increase the chance of identifying events which may influence the market. The proposed method adds further documents with a self-training strategy.

The proposed method has a clear advantage over the competing methods when evaluated by estimated F-Measure. The trading evaluation is less clear. The proposed strategy obtains the single highest return, which was the model induced from headlines. The proposed methodology performs less well with: majority voting, description and story text. The rule trained classifier return the highest return (or lowest loss) for majority voting and description, however the alignment strategy returned the best returns for the story text classifier. It is reasonable to suggest that the two evaluation tasks indicate that the proposed method has a demonstrable superiority to the competing methods if the correct financial instrument for alignment is selected.

8.1. *Future Work*

As stated earlier the trading evaluation task is not a realistic simulation of a trading environment because the competing techniques have access to all of the news published on a single trading day. Traders have access to news published up to the time of their trade. An immediate aim of the authors will be to evaluate the competing techniques with news published when the market is closed (over-nights). The successful prediction of the direction of the opening price will be a more realistic simulation. A secondary aim will be to assign a relevance measure to each news story. The assumption that each classified story has an equal effect is arguably a naive approach because certain stories will be more relevant and therefore have a stronger effect on the market index. For example, UK unemployment figures will have a stronger effect on the FTSE-100 than the NASDAQ (an index located in the U.S.A) because unemployment in the UK is more relevant to the FTSE-100. A third priority will be to utilize news volume. The current work provides an evaluation baseline for news story classification.

References

S. Abney. *Semisupervised Learning for Computational Linguistics*, chapter Self-Training and Co-Training. Chapman and Hall, 2008.

A. Alias. Lingpipe 4.1.0. http://alias-i.com/lingpipe, 2008.

J. Almeida and U. Pinto. Jspell – um mdulo para anlise lxica genrica de linguagem natural. In *Actas do X Encontro da Associao Portuguesa de Lingustica*, pages 1–15, vora 1994, 1995.

A. Balahur et al. Sentiment analysis in the news. In N. Calzolari, editor, *Proceedings of LREC*, Valletta, Malta, may 2010.

C. Barker. Continuations in natural language. In *Fourth ACM SIGPLAN Continuations Workshop (CW'04)*, 2004.

F. Benamara et al. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of ICWSM*, 2007.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0.

W. Chan. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.

H. Cunningham et al. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the ACL*, 2002.

A. Esuli and F. Sebastiani. Sentiwordnet a publicly available lexical resource for opinion mining, 2006.

C. Fellbaum. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA, 1998.

V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the ACL*, pages 174–181, 1997.

N. Indurkhya and F. Damerau. *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2010. ISBN 1420085921, 9781420085921.

A. Kloptchenko et al. Combining data and text mining techniques for analysing financial reports. *Int. Syst. in Accounting, Finance and Management*, 12(1):29–41, 2004.

V. Lavrenko et al. Language models for financial news recommendation. In *Proceedings 9th International Conference on Information and Knowledge Management*, pages 389–396. ACM Press, 2000.

B. Levin. *English verb classes and alternations: a preliminary investigation" by Beth Levin.* The University of Chicago Press, 1993.

L. Mitra and G. Mitra. *The Handbook of News Analytics in Finance*, chapter How news events impact market sentiment. Wiley Finance, 2011a.

L. Mitra and G. Mitra. *The Handbook of News Analytics in Finance*, chapter Applications of news analytics in finance. Wiley Finance, 2011b.

M. Mittermayer and F. Gerhard. Newscats: A news categorization and trading system. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 1002–1007. IEEE Computer Society, 2006. ISBN 0-7695-2701-9.

M. Mittermayer and G. Knolmaye. Text mining systems for market response to news: A survey. Technical report, University of Bern, 2006.

S. Pfeifer. Why defence should prove to be defensive, 2009. http://goo.gl/elq71 consulted in 2009.

V. Subrahmanian and D. Reforgiato. Ava: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems*, 23(4):43–50, 2008. ISSN 1541-1672.

N. Taleb and A. Lane. *The Black Swan (The impact of the highly improbable.* Random House, 2008.

J. Thomas. *News and Trading Rules.* PhD thesis, CMU, 2003. http://goo.gl/4NVHa.

D. Wood, editor. *Linking Enterprise Data*, volume 1st Edition, chapter The Role of Community-Driven Data Curation for Enterprises. Springer, 11 2010.

B. Wuthrich et al. Daily prediction of major stock indices from textual www data. In *Proceedings of KDD-98*, pages 364–368, 1998.