# Identifying polarity in financial texts for sentiment analysis: a corpus-based approach

**Antonio Moreno-Ortiz**

**Javier Fernández-Cruz**

UNIVERSIDAD DE MALAGA

Tecnolengua

*Tecnologías Lingüísticas y Comunicación Intercultural - Grupo PAI en S.I.C.A.: TIC219*

http://tecnolengua.uma.es

# Table of Contents

# Introduction

Most SA systems are designed for specialized domains using domain-specific corpora as training data for **machine learning algorithms** that classify an input text as either positive or negative.

# Introduction

A few SA systems are **lexicon-based**, where sentiment-carrying words and phrases are collected and then searched for during analysis, to come up with a certain sentiment index.

# Introduction

ML SA systems work well for specialized domains, but not so well for general language. Need to be trained for each domain.

Lexicon-based SA systems are hard to develop: high quality lexical resources are key to good performance. Acquisition process is hard. Domain specificity is an issue.

# 2. Lexicon-based Sentiment Analysis

## Sentitext & Lingmotif

Developed by the Tecnolengua group (UMA).

**Sentitext**: lexicon-based SA system for Spanish with sentence-level contextual valence shifters. Originally designed for general language.

**Lingmotif**: Integrates English. Detects languages automatically. Aims to provide a solution to the domain-specificity issue.

# Domain-dependent SA

The limitations of Sentitext for domain-specific texts were evidenced and discussed in previous work:

A set of hotel reviews from the Tripadvisor website were analyzed:

- Certain qualities **keep the orientation** they show in general-language texts, e.g., cleanliness (of rooms), character (of staff)
- Others acquire a **specific orientation**, e.g., location, size (of rooms, beds, etc.)

# Analysis process

1. Basic NLP tasks: preprocessing, tokenization, language detection, lemmatization, POS tagging.

2. MWE identification and and tagging.

3. Lexical words and MWE are looked up in the sentiment lexicons. If found, they are assigned the corresponding valence.

4. Context rules are searched for each lexical word/MWE. Matching segments are assigned the valence resulting from the application of the context rule.

5. Affect intensity (i.e., the proportion of sentiment-carrying vs. neutral units) is calculated.

6. The final Global Sentiment Value (GSV) and star rating is calculated.

# Data Sources

**Lingmotif** uses three major linguistic data sources for each language:

1. The individual words dictionary: 14K words

2. The multiword expressions dictionary: 24K entries (still growing!)

3. The context rules set: our implementation of CVS

Lexical items were assigned a valence marking their orientation and degree (from -2 to 2).

Contains neutral-polarity items, which are needed in order to block polarity-laden words that are part of them.

# Context Rules

CRs can modify a lexical item by **intensifying** or **inverting sentiment** i.e.:

1. [be] a total + neg adj: “She’s a total loser”

2. [be] no + neg adj: “He’s no fool”

# Domain-dependent SA

We could tweak the lexical resources: for example by introducing certain recurrent phrases as MWE (e.g., “small beds”), but this cannot account for all subject domain issues (and there are many domains!)

Our solution: acquire specialized lexical resources for subject domains, implemented as **plug-ins** that can be used optionally.

Users select which domain plug-in to use.

Text classification techniques could be integrated for automatization.

# 2. Lexicon-based Sentiment Analysis

## Reducing complexity

We are not interested in extracting all terms, BUT:

- only those that indicate positivity or negativity within the particular domain, AND
- only when their orientation differs from the one they exhibit in general language or other domains.

SO

- *analyst*, *sale* and *investor* are irrelevant to us, since they are neutral
- *recovery* and *unemployment* are always negative, and they are already accounted for.

# 3. Method

3-step method:

1. Extract candidate terms from specialized corpora.
2. Match terms against our general-language polarity database.
3. Obtain domain-specific polarity sentiment-bearing words.

# 3. Method

## STEP 1

Term extraction:

Gillam & Ahmad's three-step algorithm based on the *weirdness ratio* (R) measure.

**Specific-domain words are identified by a higher weirdness ratio** in the specialized corpus than the one they have in general language corpus.

# 3. Method

## Corpora:

- Specialized language: “Mag-Finance” and “News-Money” sections of the Corpus of Contemporary American English (COCA), 7.97M words.
- General language: Corpus of Global Web-Based English (GloWbE), 1,9b words.

## Samples:

2 financial news short texts (Reuters) from January 2105:

- Sample A. Types: 210 Tokens: 384 TTR: 0.546875
- Sample B: Types: 233 Tokens: 475 TTR: 0.490526

# 3. Method

## STEP 2

Manual procedure to identify relevant terms::

1. Check the semantic orientation of each candidate term by analyzing them in context.
2. Discard neutral terms, i.e., those whose meaning does not convey any particular semantic orientation.
3. Match the list of polarized terms against our existing list of polarized words.
4. Discard terms whose polarity matches (both in orientation and intensity) our existing general-language words.
5. The remaining terms are candidates for the specialized lexicon items.

# 3. Method

## STEP 3

1. Use intuition to **handpick** those words likely to convey any some semantic orientation:

- *analyst, bond, exports*: clearly neutral
- *crisis, debt*, *inflation*: clearly polarity words.

2. Check speculation against text data (**concordancing**)

# 4. Results

After STEP 1:

**98 term candidates** from two sample texts.

- 83 true positives
- 15 false positives
- 24 false negatives
- 125 true negatives

**Precision: 84.69%**

**Recall: 77.57%**

**Accuracy: 84.21%**

# 4. Results

After STEPS 2 and 3:

Candidate terms were matched against our general language lexicon.

Possible results and actions taken:

- Term does not exist in the general-language lexicon → add it to specialized sentiment lexicon: *bailout, boost, carryover, expand, expansion, flat, shrink, slow, slump, tight, wane*
- Term is present in the general-language lexicon, but the orientation varies → add it to specialized sentiment lexicon (no cases were identified)
- The candidate term is present in the general-language lexicon with the same orientation → rejected as redundant: *debt, decline, recovery, reform, unemployment*

# 4. Results

## True Positives (83)

analyst
**bailout [-]**
bank (n)
bond
**boost [+]**
**bubble [-]**
capital
**carryover [-]**
**cautious [-]**
**charge (v) [-]**
commission
**consolidation [+]**
construction
contract (v)
**crisis [-]**
current
**debt [-]**
**decline (n) [-]**
demand
domestic
**drop (v) [+/-]**
economics
economist
economy
estimate (n)
estimate (v)
**expand (v) [+]**
**expansion [+]**
expect (v)
exports
**fall (v) [+/-]**
finance
financial
**flat (j) [-]**
forecast (v)
**fuel (v) [+]**
**gain [+]**
gross
**grow [+]**
**growth [+]**
hold
import (n)
inch (v)
**increase [+/-]**
industry
**inflation [-]**
investment
investor (n)
job (n)
**jobless [-]**
long-term
market (n)
**meager [+]**
net (j)
office
percent
president
private
product
quarter
rate (n)
recent
**recession [-]**
**recovery [+]**
**reduction [+/-]**
**reform (v) [+]**
**rise [+/-]**
sale
service (v)
share (n)
**shrink [-]**
signal (v)
**slow [-]**
**slump [-]**
spending (n)
take (v)
**tight [-]**
**unemployment [-]**
vice (n)
**wane [-]**
workforce
year
yield (v)

# 4. Results

## False Positives (15)

**bleak [-]**
bumper (j)
largely
last
million
month
new
pace
prompt (v)
push (v)
say (v)
stem (v)
talk (v)
turnaround
union

# 4. Results

- **"Bailout":** this word has evolved to be unequivocally associated with financial problems.
- ***Expand*** and ***expansion****:* can have any polarity in general language: always positive in finance.
- ***Grow*** *and* ***growth*** and ***tight****:* appear to exhibit exactly the same behavior.
- ***Carryover****:* "a quality passed on from a previous situation": although in other domains the transfer can be of any kind, in finance, the transferred quality appears to be always negative.

# 5. Conclusions

- Combining statistical term extraction, semi-automatic filtering and concordancing, appears to be well-balanced in terms of cost-effectiveness
- Even using such as small sample, we have been able to identify a fair number of domain-dependent affect-laden lexical items and context rules

# 5. Conclusions

Our work offers practical results for the acquisition of our lexical resources, BUT

we have also been able to gain insight into specialized languages from a perspective that usually gets no attention:

**emotion permeates language, and languages, both general and specialized**

**maybe we should pay more attention to the evaluative aspect of language**

# Thanks!


Tecnolengua
Tecnologías Lingüísticas y Comunicación Intercultural -
Grupo PAI en S.I.C.A.: TIC219