The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs

Thomas Gaillat, Manel Zarrouk, André Freitas, Brian Davis
Insight Centre for Data Analytics NUIG, Insight Centre for Data Analytics NUIG, University of Passau, Maynooth University
Galway Ireland, Galway Ireland, Passau Germany, Maynooth Ireland
{thomas.gaillat, manel.zarrouk}@insight-centre.org
andre.freitas@uni-passau.de
brian.davis@mu.ie

Abstract

This paper introduces the three SSIX corpora for sentiment analysis. These corpora address the need to provide annotated data for supervised learning methods. They focus on stock-market related messages extracted from two financial microblog platforms, i.e., StockTwits and Twitter. In total they include 2,886 messages with opinion targets. These messages are provided with polarity annotation set on a continuous scale by three or four experts in each language. The annotation information identifies the targets with a sentiment score. The annotation process includes manual annotation verified and consolidated by financial experts. The creation of the annotated corpora took into account principled sampling strategies as well as inter-annotator agreement before consolidation in order to maximize data quality.

Keywords: Sentiment Analysis, Opinion, Corpus, Finance, Stock-market, Microblogs, Polarity Annotation

1. Introduction

The objective of this paper is to report on the creation of three Gold Standard (GS) corpora in the domain of Sentiment Analysis (SA) related to financial microblogs. The purpose is to create three annotated corpora which support the quantification of the polarity of an opinion with regard to target named entities (e.g. a stock or a company). The produced gold standards represent sentiment values in a fine-grained fashion: (i) by identifying the polarity in relation to specific target entities and (ii) by using a continuous polarity scale. The corpora are in English, German and Spanish.

Today's ubiquity of social media services in people's lives has lead to the daily posting of vast amounts of data from users in many languages. Many of these postings bear a common characteristic that is of holding judgment. Compiling opinions has become a major stake for organisations and people seeking to have a clear view of opinions regarding specific entities of interest. In a highly competitive and volatile domain such as finance, acquiring an accurate vision of operators' opinions at scale may give a competitive edge in sell-and-buy decisions. Microblogs posted on social media such as Twitter [1] or StockTwits[2] are a central data source to understand the market perception.

Research in the area of Sentiment Analysis (SA) can help address this need but it faces two kinds of issues. Firstly, SA is domain dependent. Much effort has been spent in the field to assign sentiments to various types of texts. Initially focused on product reviews the research has moved to other domains and other types of texts such as microblogs. SA results show that different topics yield differences in SA accuracy (Liu, 2012, p. 36). While SA models for product reviews may already have achieved high classification accuracy, other types of texts might still present specific linguistic analysis and modeling challenges. In this respect, financial microblogs display a number of linguistic properties that are domain specific and that makes their analysis complex. Texts are short and include many references to concepts that have a specific terminology. Consequently, SA methods must be adapted to the syntactic and semantic features of financial microblogs. The second kind of issue for SA is that of multilingualism. Many SA methods rely on surpervised machine learning strategies. As such they require annotated data sets for model training in the target languages. Few annotated resources exist in German and Spanish in the domain of microblogs and even less in the financial domain.

To the best of our knowledge, there is no resource available publicly that is dedicated to such synthetic messages with multiple target entities and languages. To bridge the gap, we built three Gold Standard corpora dedicated to financial microblogs. The corpora are in three different languages (English, Spanish and German) and sizes. After extraction, the data were manually annotated by several independent expert annotators. Annotation comes in the form of a score showing positive or negative gradual degrees of opinions about entities included in sentences. Possible exploitations of the corpora include tasks such as construction and evaluation of SA models and linguistic analysis of the investment domain.

The rest of this paper is organized as follows. In Section 2., we cover existing literature in annotated corpora for SA and finance. Section 3. details the data collection method put in place to ensure representative samples. Section 4. describes the annotation process and quality analysis results are presented in Section 5.. Finally, we conclude in Section 6..

2. Related work

A large body of literature exists in the domain of Sentiment Analysis. Since our problem was to identify multilingual

[1] https://twitter.com/
[2] https://stocktwits.com/

financial resources for use in supervised learning methods, we narrow our review to annotated corpora of the financial domain. Work on financial corpora has a long history, starting with the Penn Treebank corpus (Marcus et al., 1993) which includes Wall Street Journal articles in American English. More recent work includes corpora targeting various types of texts and different languages.

A number of relevant financial corpora exists in English. (O'Hare et al., 2009) analysed financial blogs. To do so, they developed a corpus of financial blogs. The annotation scheme targets 500 specific companies and is applied at document and paragraph level, not sentence level. Annotators successively used a 3 and a 2 point polarity scale of sentiment. An annotation tool was used by seven trained raters who annotated 979 documents. Inter-rater agreement shows a 0.712 Kappa for the three point scale and perfect agreement for the binary scale. The annotated corpus made up of 1,691 document-level annotations is subsequently used in a sentiment polarity classification approach. In (Malo et al., 2013), the authors trained classifiers to conduct sentence-level analysis of financial news sentiments (not microblogs). They used a human-annotated phrasebank with close to 5,000 sentences collected across a number of financial news sources. The annotation scheme is designed with a 3 point scale polarity. Three trained raters annotated the corpus and pairwise Kappa values ranged from 0.611 to 0.886. Majority vote is used to consolidate the Gold Standard. (Takala et al., 2014) focused on annotating a corpus of newswire texts categorised in ten different topics in economics. They addressed the need to have sentence level annotation to determine the aggregated sentiment for a given topic at document-level. In total, 297 texts were annotated both at document and sentence levels, including a three point scale polarity. Pairwise agreement Kappa for documents ranges from 0.412 to 0.897 and from 0.682 to 0.756 for sentence level.

In other languages, several corpora have been published. The TASS Spanish corpus focuses on microblog posts from Twitter (Román et al., 2015). It is composed of 68,017 messages in ten topics such as politics, entertainment, literature and soccer. Polarity is indicated at text and entity levels. No inter-annotator agreement results are given. The Emotiblog corpus (Boldrini et al., 2009) includes a 30,000 word Spanish component. It includes message blogs that focus on topical news such as the Kyoto protocol but not on financial topics. The fine-grained annotation scheme includes a mix of syntactic and semantic features. Average inter-annotator agreement across all categories is 0.68. In German, there are two projects of interest. Scholtz et al (Scholz et al., 2012) published a corpus of news-related texts. It includes 15,089 sentences and includes 3-level polarity scale (Cohen's Kappa = 0.88). Momtazi (Momtazi, 2012) published another German corpus. It focuses on blog messages and includes 500 short texts from social media. Messages are annotated at document level with four polarity scores. Fleiss' Kappa scores range from 0.5 to 0.65.

This short review indicates that microblog annotated corpora focused on the financial sector are scarce. In addition, very few of these resources provide polarity information at entity level. The SSIX GS corpus addresses this lack by providing polarity annotation, on a continuous scale, for financial entities mentioned in microblogs.

3. Data collection

The gathering process of the data requires the actual collection from data sources and the extraction of the data with filters to eliminate irrelevant contents such as spams and non-reliable sources.

3.1. Data source

The corpora are specialised in the domain of stock-market investments. They are intended to represent financial microblog messages in each language. A number of platforms such as Twitter and Stocktwits provide messages published by their users who are investors with a specific interest in following stock prices. They share stock trend analysis and relevant events. Our sampling frame is a collection of messages from the StockTwits and Twitter platforms for their focus on financial entities called *cashtags*, e.g., *$AAPL* for Apple's stock. The texts are classified according to the *cashtag* categories that refer to companies or stocks. The messages are in English, Spanish and German.

StockTwits and Twitter are social media platforms. StockTwits was used as a source for the English corpus. The reason is that it provided an official firehose access or on-demand historical English data specialised in the financial domain. The period of collection was between October 2011 and June 2015. The Spanish and German corpora were sourced from Twitter for several weeks between October and November 2016. They underwent a less sophisticated sampling process due to limited access on demand. The data were then passed through further content filtering operations (to exclude spam and unreliable authors) which are described below.

3.2. Extraction process

We explain the sampling process from the data sources. We then describe how the data was filtered to obtain relevant messages.

3.2.1. Sampling the data source

The extraction process was conducted in two stages. Firstly, samples were collected in each language by applying a stratified strategy. The purpose was to ensure representativeness in the corpora (Biber, 1993). We first categorised messages according to their time stamps and their *cashtags*. We then randomly sampled messages from all categories to make sure that each sample included a balanced representation of messages according to time and entities.

3.2.2. Filtering the data

After sampling was applied, a number of filtering rules were identified to eliminate non-desirable messages in each corpus. Such messages contained spams and irrelevant content. To detect irrelevant content we used the following criteria: *Number of messages per author*, *number of cashtags mentioned*, *number of message followers*. Once detected the corresponding messages were eliminated and three language-specific corpora were prepared for manual annotation.

4. Annotation

After extracting messages from the data sources, we conducted a manual annotation of the corpora.

4.1. Description of the Annotated Corpora

The corpus is divided into three corpora that match the language of the messages. Table 1 shows the properties of each corpus. The corpora include the messages and their

Corpus Language	Number of annotated messages	Number of unique entities annotated	Number of tokens
English	1,336	805	16,567
Spanish	766	182	11,482
German	784	101	13,660

Table 1: The English, Spanish and German corpora

score annotations. In each message the targeted entity is assigned a score. If a message included several entities, each entity was rated independently, resulting in identical or different scores depending on the context. The following is a Spanish message with its annotation including sentiment polarity and its target:

- `ID`: 1,069
- `Source ID`: 272
- `Text`: $BAC reportó ganancias e ingresos mejores que los proyectados, impulsados por el trading de bonos #stocks (Translation: $BAC reported higher than planned profits and revenues, which were pushed up by trading bonds #stocks)
- `Target`: $bac
- `Average score`: 0.62
- `Final score` : 0.412

The *$BAC* message target appears in the text and is assigned a positive score. The average score represents the average of the scores assigned by the annotators. The final score represents the consolidated score. The ID is the unique identifier within the corpus and the source ID is the message identifier of the microblog platform.
In terms of textual characteristics, financial microblogs demonstrate domain-related and linguistic specificities. Much domain terminology is used to describe assets, events and intentions, e.g., *calls*, *puts* indicating buy or sell options. The syntax is also specific as parataxis appears to be predominant. The following excerpts from the English GS (JSON format) show simple messages in which strong and even weak punctuation or symbols are used as a separators for clauses.


```
{
"Unique":  "14908$TZA",
"Text":  "Buying $TZA calls / $TNA puts
tomorrow",
"Average Score":  0.841,
"Type":  "stocktwits",
"source_id":  5350721
},
{
"Unique":  "14975$SPY",
"Text":  "$SPY Yes buy everything up,
World is fine",
"Average Score":  0.829,
"Type":  "stocktwits",
"source_id":  6058084
}
```


4.2. Applying the Annotation Scheme

The three corpora underwent a twofold process: annotation and consolidation.

4.2.1. Annotation

Firstly, several expert annotators annotated the messages by applying a score to each financial entity found in each message. Scoring involved assigning values along a continuous range of [-1;1]. This range covers seven categories that were detailed in specific guidelines for annotators. On a scale of 1 to 7, the annotators were asked to annotate how bearish (negative) or bullish (positive) the opinion holder was regarding a specific financial entity mentioned in a given message. The annotations spanned from clearly negative (*selling intention*) to clearly positive (*buying intention*). The intermediary categories showed different degrees on the scale, and 0 was kept for purely informative messages in nature. The annotators were instructed not to overuse neutral scores if possible.

4.2.2. Consolidation

The second part of the annotation process consisted in consolidating the different scores assigned to each message by the annotators. It implied the systematic flagging of inconsistent scores for the same message (e.g., averaging positive and negative values). We applied two filtering rules. The first one consisted in setting a minimum number of required annotation scores (set to 3) per entity in order to proceed with computing their average. The second rule conditioned the calculation to only taking scores within a specific deviation of one another. If a score was over the deviation, it was not included in the consolidated score. These two rules ensured a degree of control over extreme and isolated scores and prevented them from being automatically included in the final set. Finally, the messages that conflicted with the two rules were flagged and reviewed by an expert for final score assignment. During consolidation, the selection of relevant text spans corresponding to scores was abandoned due to the lack of agreement between annotators.

5. Quality tests and results

Different measures were applied at different stages in the process to evaluate the quality of the corpora. First, we applied a binomial test (Baroni and Evert, 2009) on the StockTwits English corpus to measure the quality of the sampling method regarding representativeness. We could not apply this to the Spanish and German corpora as, unlike StockTwits, we could not access the full population for the time

period due to the sampling restrictions of the Twitter API. Results showed that a low number of tokens were proportionally different between the initial data set and the corpora (see Table 2).

Total number of tokens (%)	72,263 (100%)
Number of tokens significantly different	1,330 (1.8%)
Number of tokens not significantly different	70,933 (98.2%)

Table 2: Number of tokens that are different and similar between the initial data source and the English corpus (p-value < 0.05)

Following Artstein and Poesio (2008), inter-annotator agreement was also measured with Fleiss' Kappa and Krippendorff's Alpha. The annotations were split into 2 classes (negative and positive). Results for each corpus are presented in Table 3.

Languages	English	Spanish	German
Raters	4	3	3
Krippendorff's Alpha	0.61	0.504	0.594
Fleiss' Kappa	0.69	0.547	0.703

Table 3: Inter-annotator agreement for the three corpora

The rates between corpora are not identical; especially with the Spanish corpus. These differences suggest that, in spite of identical annotation guidelines, Spanish-speaking annotators deviated from their English and German-speaking counterparts. This may indicate variations in interpreting the degrees of sentiments in the examples given in the English guidelines.

Overall, agreement results are comparable to the state-of-the-art annotation experiments mentioned in Section 2.. Interpreting the values is a matter of discussion in itself as there is not any definite scale that expresses exact agreement levels, in particular under a continuous scale setting. For a recent interpretation on consensus analysis the reader is referred to (Artstein and Poesio, 2008).

6. Conclusion

In this paper, we have presented the compilation and annotation of three gold standard corpora for financial microblogs. We followed a twofold process. Firstly, we extracted the data from the Stocktwits and Twitter microblog services. The messages were filtered to ensure the elimination of irrelevant messages such as spams and bot-authored texts. Secondly, we created the annotated corpora by applying a sentiment polarity scale at entity-level in three different languages, i.e., English, Spanish and German. These corpora were annotated by following strict guidelines by several experts in the financial domain.

The corpus data structures follows the TSV data-exchange format, the English corpus is available online[3] and the Spanish and German corpora are available on request. Future work will involve the integration of more languages in the corpus.

7. Acknowledgements

We would like to thank all the people involved in the creation of the Gold Standard. This work is funded by the SSIX Horizon 2020 project (Grant agreement No 645425)

Co-financed by the European Union
Connecting Europe Facility

8. Bibliographical References

Artstein, R. and Poesio, M. (2008). Inter-coder Agreement for Computational Linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Baroni, M. and Evert, S. (2009). Statistical methods for corpus exploitation. In Anke Lüdeling et al., editors, *Corpus Linguistics An International Handbook*, volume 2, pages 777–803. De Gruyter, Berlin, Boston.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Boldrini, E., Balahur Dobrescu, A., Martínez-Barco, P., and Montoyo, A. (2009). EmotiBlog: a fine-grained model for emotion detection in non-traditional textual genres. In *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis*, pages 22–31, Seville, Spain. WOMSA.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, San Rafael, Calif., May.

Malo, P., Sinha, A., Takala, P., Ahlgren, O., and Lappalainen, I. (2013). Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, ICDMW '13, pages 945–954, Washington, DC, USA. IEEE Computer Society.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Momtazi, S. (2012). Fine-grained German Sentiment Analysis on Social Media. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association (ELRA).

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., and Smeaton, A. F. (2009). Topic-dependent Sentiment Analysis of Financial Blogs. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 9–16, New York, NY, USA. ACM.

[3] https://ssix-project.eu/knowledgebase/gold-standards/

Román, J. V., Cámara, E. M., Morera, J. G., and Zafra, S. M. J. (2015). TASS 2014 - The Challenge of Aspect-based Sentiment Analysis. *Procesamiento del Lenguaje Natural*, 54(0):61–68.

Scholz, T., Conrad, S., and Hillekamps, L. (2012). Opinion Mining on a German Corpus of a Media Response Analysis. In *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 39–46. Springer, Berlin, Heidelberg, September.

Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2152–2157, Reykjavik, Iceland.