CoSACT: A Collaborative Tool for Fine-Grained Sentiment Annotation and Consolidation of Text

Tobias Daudert[1*] , **Manel Zarrouk**[1] and **Brian Davis**[2]

[1]Insight Centre for Data Analytics, National University of Ireland, Galway
[2]Department of Computer Science, Maynooth University
{tobias.daudert, manel.zarrouk, brian.davis}@insight-centre.org

Abstract

Recently, machine learning and in particular deep neural network approaches have become progressively popular, playing an increasingly important role in the financial domain. This results in an increased demand for large volumes of high quality labeled data for training and testing. While annotation tools are available to support text analysis tasks such as entity recognition and sentiment classification for generic content, there is an absence of annotation tools purposely built for the financial domain. Hand in hand with this, there are also no existing best practices for sentiment annotation in the financial domain. To address this issue, we suggest fundamental practices for the creation of new datasets for this domain and integrate these into our annotation tool. We present CoSACT, a server-based tool which supports the collaborative annotation and consolidation of a dataset purposely built for the financial domain.

1 Introduction

The increase in popularity and adoption of social media in recent years has generated a deluge of valuable, high volume and volatile user-generated content [Derczynski *et al.*, 2015]. Microblogs have become a central medium of communication and are used in almost all areas of our daily life. They are mainly characterized by their short length, nonetheless, they can be rich in content and highly opinionated [Sinha, 2014]. Microblogs are used by services such as Stocktwits[1], Twitter[2], or Facebook[3]. The growth of Twitter, by 1,006% between 2010 and 2015, clearly shows the high importance of short messages.[4]. Opinion Mining from short texts, such as microblogs, has become an increasingly important task [Derczynski *et al.*, 2015; Derczynski *et al.*, 2013]. Data generated from online communication acts as a potential goldmine for discovering knowledge [Dey and Haque,

---

[*]Contact Author
[1]https://stocktwits.com/
[2]https://twitter.com/
[3]https://www.facebook.com/
[4]https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

2009]. In the area of finance, sentiment acquired from text can positively influence businesses in many ways. For example, it can be used as an indicator for trading or provide knowledge about the public perception of a company. Sentiment classification and opinion mining tools increasingly employ machine learning approaches [Missen *et al.*, 2013; Pang and Lee, 2008] requiring access to labeled data for training and testing for benchmarking learning systems. The annotation and subsequent consolidation of microblogs into a high-quality gold standard is a challenging, time-consuming and labor-intensive task. Annotation tools aim to reduce the burden associated with this task, however, there are few publicly available annotation tools dedicated to capturing the sentiment in short texts [Trakultaweekoon and Klaithin, 2016] and often, unlike our tool, they do not focus deeper beyond the "document level" towards fine-grained annotation of sentiment annotation at the entity/target level.

In this paper, we present CoSACT[5], the first annotation tool incorporating a consolidation mode purposely built for the financial domain.

2 Annotation Design

Sentiment in the financial area is driven by multiple factors, hence, the quality and information content of datasets is crucial to the success of sentiment analysis. The results of sentiment analysis can have significant implications in the financial domain, positively as well as negatively. One example is algorithmic trading; on one hand, it can be very profitable and some companies (e.g. Renaissance Technologies[6]) are very successful with the computational analysis, on the other hand, algorithmic trading can lead to big losses when computers misinterpret information. Incidents such as the Lululemon Athletica Inc. case highlight this risk; their stocks dropped due to problems with the company's product - sport trousers - which was sarcastically discussed on Twitter. Automated trading systems misinterpreted these tweets as positive and long-investors lost money[7]. The more diverse information is available with a high level of detail, the better a classifier

---

[5]https://github.com/TDaudert/CoSACT
[6]https://www.rentec.com
[7]https://www.theepochtimes.com/how-hedge-funds-use-twitter-to-gain-an-edge-in-trading_2227349.html

is able to learn the underlying data characteristics. Prior to developing our annotation tool, we considered which short financial text parameters are necessary to conduct high-quality sentiment analysis; hence, our tool was developed keeping this in mind. We detail these characteristics below:

1. **Sentiment granularity** - Currently, the majority of sentiment datasets is annotated in a categorical fashion with polarity (positive, negative, neutral) [Saif *et al.*, 2013]. Although this simplifies the annotation process and leads to higher levels of inter-annotator agreements, a sentiment classifier trained on a polarity dataset will not be able to learn how to classify data with a higher granularity (i.e. in more than 3 classes). However, in the financial domain, this poses a severe drawback since the ability to capture the degree to which one text is more positive/negative than another is highly desirable. Considering a case in which one text is strongly positive and another one slightly negative, the question arises which text has a stronger impact on an asset price. This case applies to all periods in which multiple data with different sentiments exist. Comparing the sentiment polarity (positive, negative, neutral) with trading actions (buy, sell, hold), polarity classification seems to be an obvious choice, however, in the real world, there is seldom a scenario with only one data artifact to be classified. To address this, data annotated with CoSACT receives a continuous sentiment score between -1 and +1 with 0 as neutral (e.g. 0.472). However the tool has the ability to later divide the annotated data categorically into classes without limiting the granularity at the time of annotation.

2. **Multi-entity annotation** - Sentiment at document-level may not be sufficiently accurate when dealing with scenarios requiring the consideration of single entities. Many documents (i.e. twits, tweets, posts) mention more than one entity, hence, a good dataset should aim to include one sentiment annotation for each contained entity. The benefit here is at hand; some texts might contain positive sentiment towards some entities and neutral or negative towards others. To accommodate this, each cashtag in CoSACT achieves a sentiment annotation. As you can see in Figure 1, the displayed text contains two entities represented by $DD and $MON; both receive an individual sentiment score.

3. **Spans** - In addition to specifying a sentiment for each entity, CoSACT also requests the span in which this sentiment is contained. This feature provides high-quality information for a sentiment classifier since this fine-grained annotation allow us to aim for an entity-level classification (i.e. determine one sentiment score for each entity in a text).

4. **Contextual dependency** - A contextual dependency is represented by additional information (i.e. image, url) apart from the text the sentiment is referring to. From the sentiment analysis point of view, this is relevant as not all information a sentiment is based on can be retrieved solely from the annotated text. A more sophisticated sentiment classifier may wish to leverage the additional external contextual information.

5. **Justification** - To facilitate the consolidation, CoSACT collects the annotator's reasoning behind his/her annotation. Moreover, the automatic evaluation of the justifications can be used for the generation of a confidence score which provides further information to a sentiment classifier. It becomes possible to give more importance to some annotations than to those the annotator was unsure of his judgment.

3 The Tool: CoSACT

CoSACT is a server-based tool, which requires only a browser (e.g. Chrome[8], Firefox[9]). It is developed using JavaScript[10] (particularly NodeJS)[11], a SQL database[12], HTML[13] and CSS[14]. As it is utilizing a server-client architecture, multiple people can collaborate online in the annotation of the same dataset, at any point in time, without requiring the storage of large amounts of data on their local machines. It also gives the flexibility to outsource the annotation task as only a browser is required and the annotators will never come in contact with the raw dataset itself. The access to the tool and, hence, the data, is protected by a user-name and password combination. Therefore, the tool's availability on the web carries a reduced security risk.

CoSACT evolved from two separate tools, an annotation and a consolidation tool, which were tested in real-world scenarios. These were employed in the Social Sentiment Indexes Project (SSIX)[15] to create gold standards in different languages to monitor sentiment related to the financial markets. Additionally, the data used in the Semeval 2017 Task 5 competition on *Fine-Grained Sentiment Analysis on Financial Microblogs and News* was generated with this annotation and consolidation tool [Cortis *et al.*, 2017]. The gold standard dataset used by Byrne et al. [2016] to predict the outcome of the Brexit poll based on tweets was also created using CoSACT.

To better detail the tool's interface, we will use numbers enclosed in parenthesis, e.g. (1), which point to specific locations in the following figures.

3.1 The Annotation Mode

The CoSACT interface for the annotation mode is shown in Figure 1. The microblog text as presented to the annotator is represented in box (3). Right to it, (4) highlights the different submission options. Here, the annotator can choose between submitting his/her annotation (green button), labeling the presented microblog as spam (red button), labeling it as irrelevant (blue button), or the annotator can click on one of

---

[8]https://www.google.com/chrome/index.html

[9]https://www.mozilla.org/de/firefox/

[10]https://en.wikipedia.org/wiki/JavaScript

[11]https://nodejs.org/en/

[12]https://en.wikipedia.org/wiki/SQL

[13]https://en.wikipedia.org/wiki/HTML

[14]https://en.wikipedia.org/wiki/Cascading_Style_Sheets
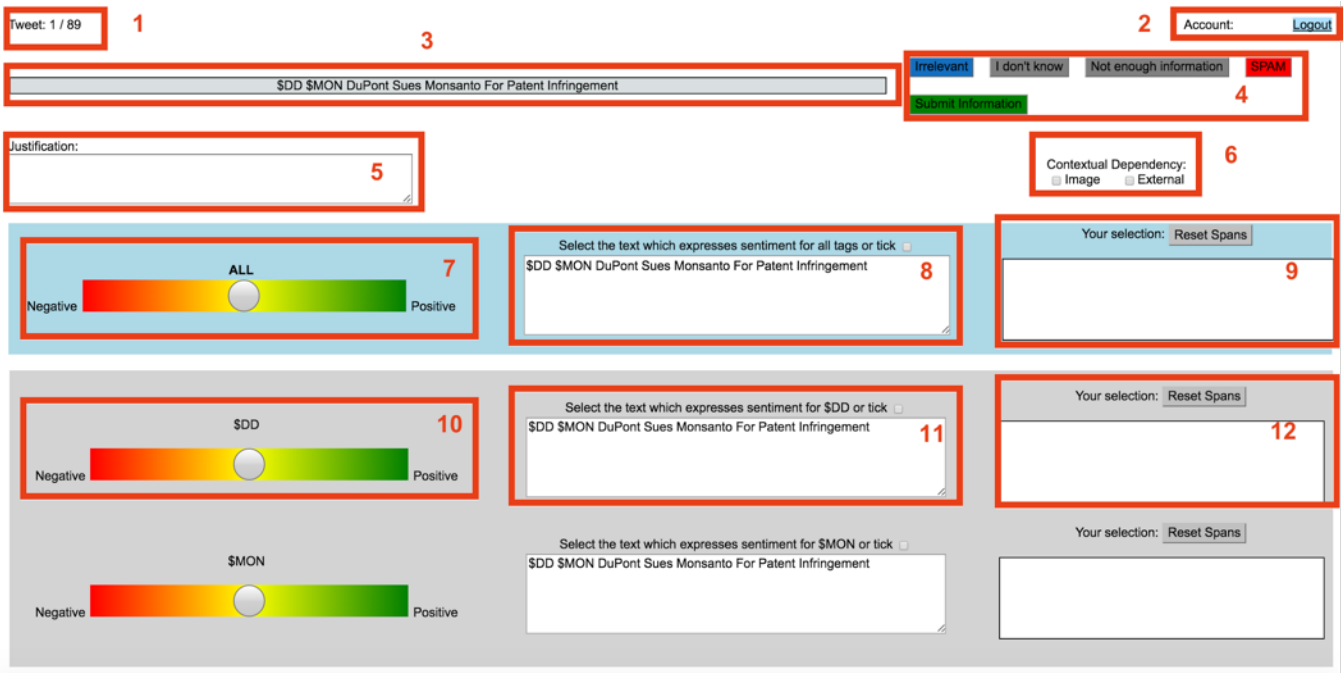
[15]https://ssix-project.eu/

Figure 1: The interface of CoSACT in the annotation mode.

the two grey buttons. In this case, the annotator then chooses to label the data as "I don't know", in case he/she is unsure about the correct annotation, or as "Not enough information", in case the uncertainty clearly derives from the presented data itself. This differentiation provides insights about the information content of the data as it becomes clear whether a text might have been too concise for a good interpretation or the annotator has simply not been able to interpret it. Below the presented tweet, (5) labels a field aiming at storing a justification for the annotators choice. This justification might provide relevant value in addition to the previously assigned feature. On its right, the two boxes (6) "Image" and "External" (e.g. URL) can be ticked in case the presented message is based on additional information and the content alone does not represent the sentiment fully. Location (10) in Figure 1 is the sentiment slider which the annotators use to assign a continuous sentiment between -1 (Negative) and +1 (Positive) to an entity. The text area at location (11), allows the annotators to assign one or multiple spans to the sentiment and, therefore, to specify the part of the microblog which contains the sentiment. In case an annotator is unable to assign a span, the box on top of the text area can be ticked. Finally, the the text area (12) in Figure 1 presents the selected spans; the button above is for resetting spans in case an annotator wants to revise his/her selection.

Note that the master selectors presented with a cyan background in (7)-(9) of Figure 1, affect all the other fields. For example, a change in the slider (7) is also moving the two sliders below. This functionality gives the annotators the capability to select spans for microblogs with multiple entities without the necessity to select them independently.

We chose to use a slider to determine the sentiment (See (7)/(10) in Figure 1) to prevent the annotators from using some sentiments disproportionately. If the sentiment range would present classes, annotators would tend to move the slider towards one of the categories (e.g. 1,2,3). Similarly, annotators requested to give numbers for a sentiment might be unable to go into a high level of detail and would tend to annotate with rounded numbers (e.g. -1.0, -0.5, 0.0, 0.5, 1.0).

3.2 The Consolidation Mode

In consolidation mode, CoSACT is able to smoothly consolidate a previously annotated dataset. At the first time the user tasked with consolidation logs in, he/she is requested to specify the number of required annotations per microblog as well as to set an acceptable deviation. The deviation defines the degree of variation between the annotations of a text. These two values are then used to automatically consolidate all microblogs which fulfill the given criteria (Figure 2). A manual consolidation is required for the remaining microblogs; automatic consolidations can be amended by the consolidator. As shown in Figure 3, the example does not meet the criteria thus, it needs to be consolidated. In box 5, microblogs to be consolidated have a gray background, auto-consolidated ones are highlighted cyan, and consolidated microblogs are shown in blue. The microblog currently in use is underlaid in green. The consolidator can use the three buttons in (4) to either classify a microblog as spam, irrelevant or to save his/her consolidation. Location (6) shows the scores from the annotators. The "ALL" area (7)-(8) is similar to areas (7)-(9) of the annotation mode, as well as the mechanism of assigning a sentiment and span. The mechanism of assigning a sentiment and spans is similar to the process described in Section 3.1. The consolidator is requested to set the final sentiment as well
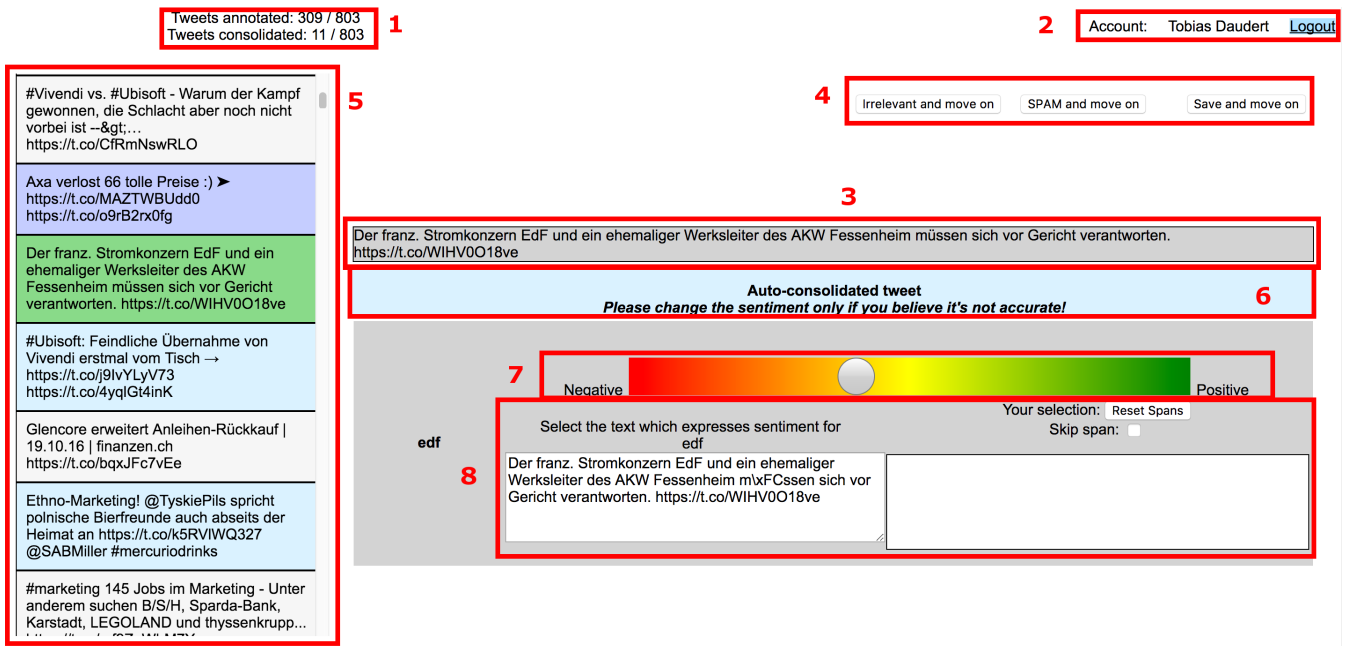
Figure 2: The interface of CoSACT in the consolidation mode presenting an auto-consolidated microblog.

Figure 3: The interface of CoSACT in the consolidation mode presenting a microblog to be consolidated.

as to set the spans, all under consideration of the annotations given and shown.

4 Related Work

Due to the current high interest in microblogs and sentiment analysis, tools focusing on sentiment annotation of short text have become of relevant importance. Some examples of ser-

vice based general purpose collaborative linguistic and semantic annotation tools include WebAnno [de Castilho *et al.*, 2014], GATE Teamware [Bontcheva *et al.*, 2013] which is integrated into the AnnoMarket Text Mining Service [Tablan *et al.*, 2013], TURKSENT [Eryigit *et al.*, 2013], and SenseTag [Trakultaweekoon and Klaithin, 2016]. Due to our interest in sentiment annotation, we focus on two of these publicly available tools which also focus on this particular task: TURKSENT and SenseTag. However, it is important to clarify that these tools are not purposely built for the financial domain nor were applied in any financial context. We chose to address these, as they are the most closely related tools to our work.

Eryigit [2013] were the first to create a publicly available sentiment annotation tool. It incorporates a manual or semiautomatic linguistic annotation layer which includes text normalization, named entity recognition, morphology, and syntax tasks. TURKSENT provides a software as a service which does not require a specific platform, and it is also accessible by multiple users. For the annotation, categorical labels supported by images (i.e. emoticons) are used; these range from 1-3 and include an additional ambivalent label. Two tasks are assigned during the annotation process. In the first task, the annotator is required to provide a general sentiment label for the text and select an appropriate comment category and sentence type. For the second step, it is performed a target based annotation where tuples of the brand, product/service and features are provided. The second analysed tool, SenseTag, consists of three components: data collection, data annotation and administrative operation. The data is automatically collected from microblogs and classified into four domains. This information is then pre-processed and stored in a database. Non-annotated messages are selected randomly and are then manually tagged for each word. The words are classified into 4 categories: positive, negative, feature, and entity. Finally, the last component allows the management of the tags, and the tagging tool as well as data handling.

Considering these tools, we identified generic tool characteristics and shortcomings, which we address with CoSACT:

1. **Complex pre-processing**. While these tools rely on an additional annotator task to deal with entity recognition, CoSACT applies a simplistic approach focusing on a previous input of entities or regular expressions (i.e. regex) to extract them. Either the entities to be found are being specified beforehand, or a regex is given to identify them. This removes the additional laborious task for the annotators. Our tool accepts free text, hence, not demanding the pre-identification of categories or features.

2. **Categorical sentiment**. Both TURKSENSE and SenseTag utilize a small, categorical range of sentiment. In CoSACT, we use a wide, continuous span from -1 to +1, which allows for the extraction of a fine-grained sentiment score. CoSACT is the only tool providing a fine-grained sentiment scale.

3. **Overall score**. Our tool provides a sentiment score for the text, as a whole, and for the entities (e.g. $Volkswagen) and/or tickers (e.g. VOW). This is not present in SenseTag.

4. **Contextual Dependencies**. Short messages may contain additional information conveyed by an image or URL. CoSACT takes this into account by allowing the user to identify messages where this situation occurs.

5. **Consolidation**. Our tool implements a consolidation mode which allows a user, preferentially a domain expert, to consolidate already annotated datasets. This is useful in cases where the minimum number of annotations is not met and/or when the deviation is exceeded, in the remaining scenarios the microblog is auto-consolidated. This is a vital step for the development of quality gold standards and not yet implemented in other tools.

6. **Usability features**. CoSACT's interface is able to indicate the number of messages in the annotation and consolidation tasks and the overall progress. It also has the advantage of easily allowing the user to review already consolidated text. Annotated text cannot be reviewed, this is a deliberate design feature. We believe the annotator should focus on the current text and not review the assigned sentiment in accordance with other messages seen before.

5 Conclusion

In this paper, we addressed the growing need for quality labeled data for training supervised sentiment classification tools for the financial domain. We established fundamental practices we believe to be essential to improve the dataset quality and incorporated them in an annotation and consolidation tool designed for the financial domain. Although collecting many different parameters from an annotator, CoSACT simplifies the process of annotation by providing an automatic consolidation as well as a manual consolidation interface. It combines the benefits of collecting a magnitude of information while keeping the manual effort low. Entities can be annotated with a sentiment, one or multiple spans expressing it, a justification for the annotation, and contextual information. In addition, data can be classified as spam or irrelevant and annotators can click on "I don't know" in case they are overwhelmed, or can mark texts with "Not enough information" in case they perceive it as not revealing enough information. All these parameters contribute meaningfully to automated sentiment analysis in a financial setting. Since this tool is server-based, it only needs to be set up once and allows annotators and consolidators to work on data with a minimum of effort, requiring not more than a standard browser. Our tool encourages the efficient engineering of quality fine-grained labeled datasets for developing, testing and benchmarking (semi-)supervised learning systems for entity orientated sentiment analysis.

Acknowledgments

References

[Bontcheva *et al.*, 2013] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029, 2013.

[Byrne *et al.*, 2016] David Byrne, Angelo Cavallini, Ross McDermott, Manuela Hürlimann, Frederico Caroli, Malek Ben Khaled, André Freitas, Manel Zarrouk, Laurentiu Vasiliu, Brian Davis, et al. In or out? real-time monitoring of brexit sentiment on twitter. *SEMANTiCS 2016*, 2016.

[Cortis *et al.*, 2017] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, 2017.

[de Castilho *et al.*, 2014] Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych, and Seid Muhie Yimam. Webanno: a flexible, web-based annotation tool for clarin. In *Proceedings of the CLARIN Annual Conference (CAC)*, 2014.

[Derczynski *et al.*, 2013] Leon RA Derczynski, Bin Yang, and Christian S Jensen. Towards context-aware search and analysis on social media data. In *Proceedings of the 16th international conference on extending database technology*, pages 137–142. ACM, 2013.

[Derczynski *et al.*, 2015] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

[Dey and Haque, 2009] Lipika Dey and Sk Mirajul Haque. Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition*, 12(3):205–226, 2009.

[Eryigit *et al.*, 2013] Gülsen Eryigit, Fatih Samet Cetin, Meltem Yanik, Tanel Temel, and Ilyas Çiçekli. Turksent: A sentiment annotation tool for social media. In *LAW@ ACL*, pages 131–134, 2013.

[Missen *et al.*, 2013] Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. Opinion mining: reviewed from word to document level. *Social Network Analysis and Mining*, 3(1):107–125, 2013.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis Bo. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, 2008.

[Saif *et al.*, 2013] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013.

[Sinha, 2014] Nitish Sinha. Using big data in finance : Example of sentiment extraction from news articles. Technical report, Board of Governors of the Federal Reserve System (US), 2014.

[Tablan *et al.*, 2013] Valentin Tablan, Kalina Bontcheva, Ian Roberts, Hamish Cunningham, Marin Dimitrov, and AD Ontotext. Annomarket: An open cloud platform for nlp. In *ACL (Conference System Demonstrations)*, pages 19–24, 2013.

[Trakultaweekoon and Klaithin, 2016] Kanokorn Trakultaweekoon and Supon Klaithin. Sensetag: A tagging tool for constructing thai sentiment lexicon. In *Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on*, pages 1–4. IEEE, 2016.