Machine Learning Techniques for Annotations of Large Financial Text Datasets

Completed Research Full Papers

Jesse Roberts
Otto-von-Guericke-University
Magdeburg
jesse.roberts@st.ovgu.de

Matthias Volk
Otto-von-Guericke-University
Magdeburg
matthias.volk@ovgu.de

Robert Neumann
Ultra Tendency GmbH
Magdeburg
robert.neumann@ultratendency.com

Klaus Turowski
Otto-von-Guericke-University
Magdeburg
klaus.turowski@ovgu.de

Abstract

Today, a number of automated methods exist to augment predictive models with annotations when working on huge collections of unstructured texts. One such method involves the use of machine learning techniques. This work seeks to investigate the use of those techniques for annotations on UIMA - a framework used for the analysis of unstructured data. By using the design science research methodology, an annotator pipeline is created as the main artifact. It takes news and blog articles as input and extracts entities, which are then annotated with a sentiment. Concurrently a demonstration is taking place, using a set of 12,480 gold annotated documents on German car manufacturers as training data. By harnessing a multitude of validations methods and machine learning algorithms, all results are thoroughly tested and evaluated. As a result, this work provides a blueprint for research into the use of UIMA and machine learning techniques for domain constrained datasets.

Keywords

Machine Learning, Annotation, Natural Language Processing, UIMA, Financial Industry, Big Data

Introduction

In recent times, improvements in research and technologies for data management and processing have garnered interest in related fields. One of such fields is that of text mining, which has extensive applications in many areas, such as in the health, financial, and social domains (Hotho et al. 2005; Ngai and Lee 2016; Sakurai et al. 2016; Stilo and Velardi 2016). Consequently, the machinery of text mining consists of many phases starting from data collection, to the analysis and the visualization of the findings (Fayyad et al. 1996; Shearer 2000). In each of those, sophisticated processes are applied like annotations (Hotho et al. 2005) that can be explained as the process of adding value to an object in the form of labels, comments, explanation, and other remarks without changing the object in question (Dingli 2011). The challenge of annotating texts has been addressed differently across the various domains of applications. Some of the recent challenges include the unstructured format of text, the high dimensional feature set, the purse size, and also determining text syntax and semantics, such as the meaning of "*VW takes over Audi*" versus "*VW takes on Audi*". Meanwhile, in the financial domain, adding context to an entity or extracting secondary features like the parts-of-speech of the text could enhance its significance. For instance, VW and Audi can be identified as proper nouns and thus highlight which and how business entities are affected by the given statement. Similar to this, fraud detection and creditworthiness prediction are some areas in the financial domain which also benefits from those techniques (Cruz et al. 2016; Hotho et al. 2005; Nassirtoussi et al. 2015). Depending on the tool or objective, annotations can be used for text pre-processing or as a labelling feature to classify text. These techniques present their own frames of implementation which may or may not be efficient for the domain of application. An opportunity is thus presented to investigate and apply language processing frameworks, algorithms, techniques, such as machine learning, and which will be

rightly suited for annotations in the financial industry. Consequently, the following research question (RQ) shall be answered in the course of this work: *How can large and unstructured financial text document datasets, originated from various sources, be annotated through the use of different techniques?*

Hence, the work seeks to investigate the applicability of annotation techniques, such as, machine learning (ML) for the annotation of financial news data sets. After initial investigations, among the available tools and software frameworks, the Unstructured Information Management Architecture (UIMA) was chosen to facilitate the found out results, especially in form of an experimentation and evaluation. UIMA is a suite of software applications that are used in the analysis of large quantities of unstructured data sets to produce meaningful value to the end user (Ferrucci and Lally 2004b). Some examples of the outputted values include entities, relationships and patterns fitting predetermined rules (Kano et al. 2009). The consequential main objective, is to present a pipeline of annotators provided by UIMA and an algorithm which gives the best results for the financial news and blog articles. As such, in the course of this research, preprocessing and annotation techniques will be combined with ML algorithms and tested using UIMA, to determine which combinations produce a well-annotated end result. As the main source, freely accessible English speaking financial news data from reliable blogs and online news articles are used which offer archives dating back at to at least ten years. The research guideline, the design science research methodology (DSRM) as proposed by (Hevner et al. 2004) and six stepped workflow by (Peffers et al. 2007) is utilized. This meticulous process provides an artifact in response to observed problems and also forms the flow of this work that is structured as follows. While in the first section an initial motivation, the research question and the main objectives were already presented, the second section deals with the state of the art. In here, various techniques are introduced to provide a concise view of text annotation in the financial domain. Section three describes the design and implementation of the noted techniques originated from the results of the second section. In doing so the main artifact is proposed and concurrently demonstrated by harnessing a use case of news article and blogs on German car manufacturing firms. Within the subsequent section, all observations from the techniques, implemented in the design phase are shown, evaluated and analyzed to see how well they meet the research goals. Additionally to that, all results are critically discussed and the significance to the area of research is presented.

State of the Art

To find an answer to the previously formulated RQ, first, the state of the art is herein presented, using the methodology by (Mayring 2003) as well as (Webster and Watson 2002). Both served as a guide for the identification, collection, categorization and analyses of the research materials. Derived from the RQ and the main objectives, initially a set of sub-RQs (SRQ) were formulated, focusing on the annotation and related areas. Those are S*RQ1: What is the domain distribution of available research on text annotation? SRQ2: What are the prevailing methods used for text annotation? SRQ3: How are text annotations represented in the respective domains?* After performing the core of the review, a series of objectives for the design of the solution are derived. To access the literature, search terms were formulated with focus on various aspects of the research area such as: text annotation, machine learning and some specific frameworks like UIMA. The search terms were each entered into the databases: Wiley, Scopus, ScienceDirect, Springer Link and ACM Digital Library and for an increased access also Google Scholar. Furthermore different criteria were applied for the collection and the later refinement. For the acceptance of a paper, all inclusion criteria described in Table 1, had to be fulfilled, whereby as soon as one of the exclusion criteria was noticed the paper was rejected.

Inclusion Criteria	**Exclusion Criteria**
The contribution is published in a conference proceeding, journal or book.	A contribution whose conference of publication falls below the score of B.
The contribution is related to one of the formulated research question.	A contribution whose SCImago Journal Rank (SJR) in the year of publication is below 1.77. (not in the top 5 %).
The contribution is written in English.	The technique has too many constraints and exceptions.

Table 1. Inclusion and Exclusion Criteria of the Structured Literature Review

At the beginning, a total of 386 papers were identified and further reduced by using a two-stepped refinement process (Mayring 2003). In phase one, the title and abstract of papers were scrutinized.

Afterwards, in phase two, the other parts of the remaining 111 papers were examined. By the end, 41 papers were selected and 4 additionally found out through the use of the subsequent forward-backward search (Webster and Watson 2002).

Descriptive Analysis

Out of the 45 remaining papers, 7 were exclusively on UIMA. As such most of the descriptive analysis was done on the remaining 38. In answer to the SRQ1 goal, we assessed the domain of interest for each paper and categorized them. Although it was not directly addressed within the initial search terms, we found that there is a significant interest in research material aimed at the financial and business domains. We met the SRQ2 goal by identifying the text annotation methods used in each paper. Here, most of the papers used either statistical natural language processing (NLP), machine learning (ML) methods or deep learning (DL) and there was one occurrence of a rule-based (RB) approach. Table 2 shows the detailed distribution of annotation methods per domain and the number of papers. In answer to SRQ3, the annotation representation used for each domain was assessed. The noted representations were semantic (Salas-Zárate et al. 2017), temporal (Pan et al. 2011) and sentiment (Ferreira et al. 2014). Especially the latter was found in the domain of financial (FI) and business (BI), such as in (Devitt and Ahmad 2007; Ferreira et al. 2014; Hu et al. 2018; van de Kauter et al. 2015a). Sentiment annotations dwell on identifying the emotional implications expressed within the text. The sentiments are then given an annotation, based on a polarized scale, for instance distributed into positive, negative and neutral (Devitt and Ahmad 2007; Ferreira et al. 2014; Hu et al. 2018; van de Kauter et al. 2015a). The high use of sentiment annotation within the financial domain could be explained by its output, simple polarity scales are easy to understand and helpful for quick decision making, regarding the direction of movement on the future prospects. Hence, the findings of this and the previously formulated RQs are further implemented within the following course of this work.

Dom.	No.	Paper
Financial	11	**NLP:**(Chan and Chong 2017; Devitt and Ahmad 2007; Ruiz-Martínez, María, Juana et al. 2012; Salas-Zárate et al. 2017; van de Kauter et al. 2015a); **ML:** (Das et al. 2017; Ferreira et al. 2014; Krishnamoorthy 2018; O'Hare et al. 2009; Schumaker et al. 2012); **DL**:(Hu et al. 2018)
Health	5	**NLP**:(Diallo et al. 2006; Verspoor et al. 2012; Viani et al. 2018); **DL**:(Mai et al. 2018; Singh et al. 2017)
Tech	2	**NLP:**(Gao et al. 2017); **ML:**(Borg et al. 2017)
Legal	2	**ML**:(Asooja et al. 2016); **RB**:(Lesmo et al. 2009)
Business	6	**NLP**:(van de Kauter et al. 2015b); **ML**:(Cruz et al. 2016; Maas et al. 2011; Zaidan et al. 2007); **DL**:(dos Santos and Gatti 2014; Socher et al. 2013)
Agricultural	2	**NLP**:(El-Beltagy et al. 2007); **ML**:(Cui et al. 2010)
Non-specific domain	10	**NLP**:(Bontcheva and Cunningham 2011; Chakrabarti et al. 2014; Dickinson 2015; Dill et al. 2003; Pan et al. 2011; Piao et al. 2017); **ML**:(Galke et al. 2017; Spina et al. 2015); **DL**:(Albukhitan et al. 2018; Chotipant et al. 2015)
UIMA	7	(Bethard et al. 2014; Ferrucci and Lally 2004a; Gotz and Suhre 2004; Kluegl et al. 2014; Liu et al. 2012; Wachsmuth et al. 2013; Wilcock 2017)

Table 2. Categorization of the Investigated Papers

Implications and Challenges of Financial Text Annotation

Due to the primary focus on the financial domain, it was further required to highlight and consider possible shortcomings. Financial news tend to be more expressive than official metric-based reports. Besides financial or economic implications for business entities, also sentimental wordings and relations to other companies can be included. This is because the narrative language allows for sentiments to be shared, such as *"It is not looking good for company A after the U.S Securities and Exchange Commission sent them a subpoena on their claims to take the company private"*. Given the variety of a text in terms of author tone,

language usage and word semantics, a number of challenges arise in the effort to efficiently automate the process of identifying sentiments from financial news (van de Kauter et al. 2015a; van de Kauter et al. 2015b). Apart from the overly sentimental wording, some of those are: identifying implicit and explicit sentiments (van de Kauter et al. 2015a; van de Kauter et al. 2015b), identifying multiple business entities of interest in texts and the relationship between them (Ferreira et al. 2014; Njølstad et al. 2014; Wang et al. 2014), temporal limitations (Pan et al. 2011), and publisher significance (Hu et al. 2018). For instance, in context of this, an article about two competing companies, may contain bad news for one company and the opposite for another company. To overcome parts of those challenges a variation of software tools can be used. After examining the literature, four main categories were identified, namely: toolkits, frameworks, on-demand annotators and corpora. Toolkits, such as Natural Language Toolkit (NLTK) or Apaches OpenNLP, are normally modular and sometimes used for specific annotation tasks like tokenization or lemmatization (Wilcock 2017). Contrary to this, frameworks are able to handle the whole annotation process from end-to-end (Ferrucci and Lally 2004b). Additionally, those sometimes include other tools which are managed as a unit, like UIMA or GATE. On-demand annotators can be accessed online (Borg et al. 2017; Socher et al. 2013). They range from human annotators like on Amazons Mechanical Turk to fully automated ones, like the Google Cloud Console (GCC). The final category, corpora is usually a dataset consisting of word ontologies and dictionaries which provide support for text annotation (Davies 2008; Ide et al. 2010). They are usually manually generated and follow standardized principles of linguistic syntax formation, such as the Manually Annotated Sub-Corpus (MASC).

Design and Development

The systematic literature review provided beneficial insights into the current state-of-the-art and thus solutions to parts of the research problem posed earlier (cp. SRQ1-SRQ3). This includes the identification of techniques, forms of annotation representation, challenges, and tools used in annotating texts in the financial domain. Consequently, a series of solution objectives were derived from the previous results to inform the design and implementation phase that recognizes those outcomes. This includes*: (i) Use of sentiment as the annotation representation for financial news. (ii) Tackling the challenge of multiple business entities per news article. (iii) Tackling the challenge of identifying implicit and explicit sentiments.* Altogether, the main artifact of this work intends to provide a pipeline of annotators and ML algorithms facilitated by UIMA to gain insights into financial news data. Hence, in fulfilment of the third step of the DSRM, this section illustrates how the literature findings and solution objectives were translated into the solution design. Apart from the general description of the main artifact, a demonstration is conducted concurrently, as expected in step four of the DSRM. For the actual design, we adhered to parts of the Knowledge Discovery in Databases (KDD) process (Fayyad et al. 1996) and Cross Industry Standard Process for Data Mining (CRISP-DM) (Shearer 2000) that are commonly used for such kind of projects. Those are also condensed within an ontology for the classification of big data technologies provided by (Volk et al. 2018). This approach maps the general phases: data provision, preparation, analysis and visualization to specific operations and related tools. Following this, the development of the intended artifact was sequentially realized and demonstrated by an accompanied use-case of German car manufacturers. A summary in the form of a business process modelling notation (BPMN) model is depicted in Figure 1. The phases of the model are further described in more detail.

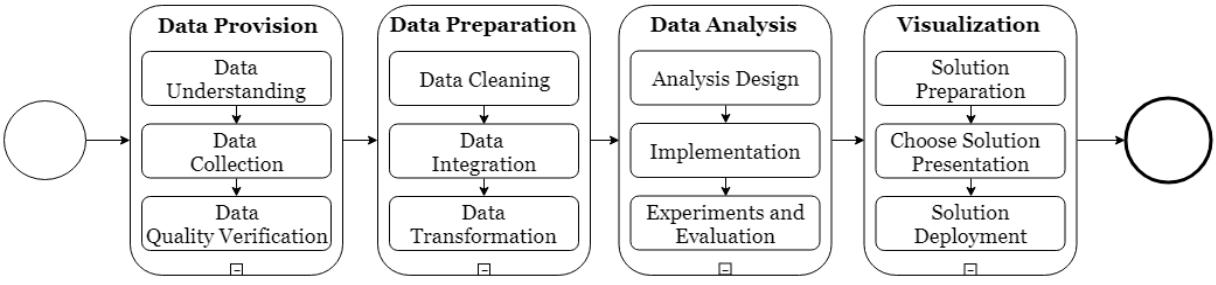

Figure 1. Condensed overview of the conducted steps in the form of a BPMN model.

Data Provision

In order to understand, collect and verify the quality of the dataset efficiently, a set of requirements influenced by the literature review were laid down. Those include, for instance, sources that hold large volume of news articles and blogs, cover an extensive range of time, originate from a valid and verifiable source and focus on financial prospects of certain entities. After this, the data was collected and its quality ensured using appropriate tools. For the actual demonstration, various strategies were considered for the initial data collection. First, authors from the conducted literature review were contacted, to receive annotated datasets originated from actual research. As a second strategy, various repositories with openly accessible data were queried. Unfortunately, none of these approaches were expedient due to failure of some of the requirements. Eventually, as the third strategy, the data was collected independently. As part of the requirements, news source sites were accepted, if news were written in English, no subscription was needed, an extensive archive containing material from more than ten years was existent and finally programmatically searchable. By the end, there were twelve news source sites out of which five (Wall Street Journal, Reuters, Spiegel, New York Times and Local.de) were selected. For demonstration, five German car manufacturers (VW, BMW, Mercedes-Benz, Audi and Porsche) were used. In all, 13,988 news articles covering a span of ten years were collected using a web scraping tool called Scrapy and stored in json format in Apache Solr. For further verification of the overall quality, the latter tool was used. Through the multitude of provided functionalities, including also text analysis features, various quality requirements originated from (Dasu and Johnson 2003; Rahm and Do 2000) were verified. In particular this applies to the completeness, consistency, accuracy and conformity of the data.

Data Preparation

In the data preparation stage, Solr was used as the main tool for the data cleaning and integration operations. The data cleaning principles of uniqueness, relevance and coherence of data entries as described by (Dasu and Johnson 2003) served as a guide in identifying duplicates and noisy data which were then removed. In doing so, the data was integrated and standardized such that each entry contained the date, headline, publisher and the article content. The resulting entries were integrated through a merge process, at which duplicating news items from different sources and irrelevant pieces, such as advertisements, were removed. Finally, by using a number of custom python scripts, the 12,480 remaining entries were transformed into a collection of documents while keeping the metadata intact in Solr.

Data Analysis

ML algorithms are trained on previously classified datasets, to classify new occurrences of similar data. Hence, in order to train algorithms, it was important to have the independently collected sample of 12,480 documents annotated. Considering the objective of handling multiple-entities and the made observations, we focused on methods that annotate each business entity within a document with a sentiment polarity of positive, mixed, negative or neutral. Given the large number of documents, we sought for an automated option and considered using GCC sentiment annotation framework. This is because of its fast output and support for entity sentiment annotation. After having the dataset annotated, the artifact – a UIMA annotator pipeline – was meticulously built iteratively and tested with various ML algorithms, such as Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM). The implementation of the algorithms were provided by ClearTK (CTK) that represents a wrapper providing ML support for UIMA. Furthermore, it also offers evaluation tools which were used to evaluate the outcome of each experiment in the pipeline. The setup of the pipeline includes a custom stop word remover using the NLTK English list of stop words, a sampler that splits the dataset into a ratio of 70 percent of training and 30 percent of testing data, a self-developed UIMA gold annotator extractor that extracts the GCC annotation, and a CTK annotator. The CTK annotator was configured to consist of a feature extraction part which extracted features such as tokens, sentences, parts-of-speech (POS) tags, named entities, proper nouns, topic terms and words near terms (WNT). The annotator includes a ML algorithm which was trained on the extracted features, a classifier that predicts a sentiment using a trained model and a separate evaluation phase. Even though k-fold cross validation requires specific attention, this was used on the training data within the evaluation phase to ensure that no overlapping's occur and each class is represented in the training phase. Additionally, the holdout validation was used for previously unseen data in order to compute annotation statistics. Therefore, the pipeline as shown in Figure 2 receives a collection of news articles as input and

creates a trained model, classification results and annotation statistics. In particular, the flow of the artifact can be described as follows. First, every document is loaded into the UIMA Common Analysis System (CAS), which is a special data type system used by UIMA for pipelines in order to maintain data format consistency. The sentence annotator then splits the CAS object into sentences while the gold annotator extracts the GCC annotations from the documents and encapsulates them as a UIMA type named entity to be used for training. After this, a token annotator splits the sentences into words, identifies words that correspond to the named entities as topic terms and extracts other words near the identified terms as WNT.

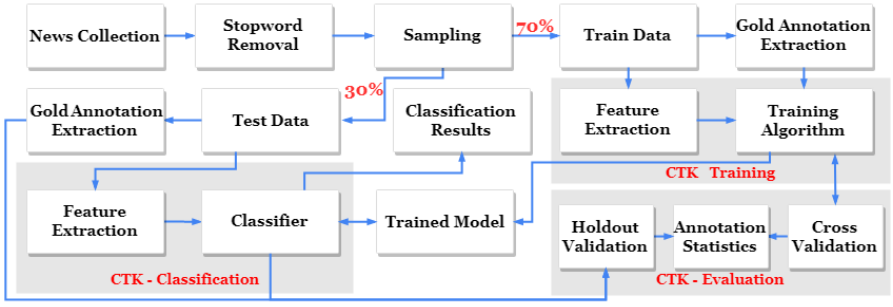

Figure 2. The Designed and Developed Artifact

Finally, the POS annotator annotates the tokens with standardized NLP tags such as "VB" for verb base and "NNP" for proper nouns to denote the part of speech of the word. All extracted features are maintained in the UIMA CAS system. After their extraction, a ML algorithm is trained on the features to get a model. Given the fact that each article was treated as having multiple entities the annotation prediction was handled as a multi-label classification problem considering multiple sentiments. A CTK evaluation library is then further used to calculate the accuracy of classifications after comparisons with the gold label annotations. The library outputs a confusion matrix containing the classification results as well as statistics metrics, like the recall, precision, F1 measure and correct classifications.

Data Visualization

The output of the pipeline was visualized to make the results of the classification and evaluation easier to understand. In the case of classification, the assigned values for each document were written to a separate text file, like "*VW (neutral), Porsche (positive)*". The confusion matrices outputted from the 10-fold cross validation and holdout validation were also written to text files, cumulated and later plotted (cp. Figure 3). The first output of entities could be used to obtain a quick insight at a glance into current public sentiment. Thus decision makers may facilitate those results to obtain the current market sentiment of a business entity. Additionally these could be used by predictive models to inspect the near future.

Evaluation

In order to demonstrate the potency of the artifact as described by (Hevner et al. 2004), relevant methods were used to evaluate the solution. This step corresponds to the fifth step of the DSRM. For each iteration of the experiments as described above, the holdout test validation and 10-fold cross validation were carried out. In here, the focus was placed on the comparison of the F1 scores, because it represents a harmonic mean of the precision (P) and recall (R). Thus it provides a more rounded understanding of the results. However, also other indicators, such as the Receiver Operating Characteristic (ROC) curve is imaginable in context of this, but are not further considered due to the nature of the data. The 10-fold cross validation was performed on the 70% of training data. This was done to make full use of the training instances without overlapping. Meanwhile, the holdout validation was carried out on the remaining 30% of the data. It was done to test how the model performs with previously unseen datasets.

Experiments

The experiments carried out involved varying some parameters of the annotators in the pipeline and measuring their influence on the results of the annotation. These parameters were the WNT and four ML algorithms, namely NB, ME, SVM and Conditional Random Fields (CRF). The WNT represents a

configurable number of words before and after an identified business entity. This parameter was chosen mainly to understand if the size of the resulting feature set had an effect on the results of the classification. CTK offered more than one library for some of the mentioned algorithms, thus multiple libraries were applied, in particular Mallet (M) and OpenNLP (O). In total, the following library-algorithm combinations were used: M-CRF, O-ME, M-NB and M-ME. For SVM the libraries LIBSVM (S) and LIBLINEAR (L) were used, which additionally allowed a modification of certain parameters like the cost (c). For each variation, a separate experiment was carried out, resulting in a total of 18 combinations. The figure below depicts the F1 score for the holdout as well as the 10-fold cross evaluation. While a two or three induces a variation of the WNT by the same, a c_4 or c_{10} refers to the application of the cost by four or ten. The results show similar patterns across all the algorithms for the F1 score.

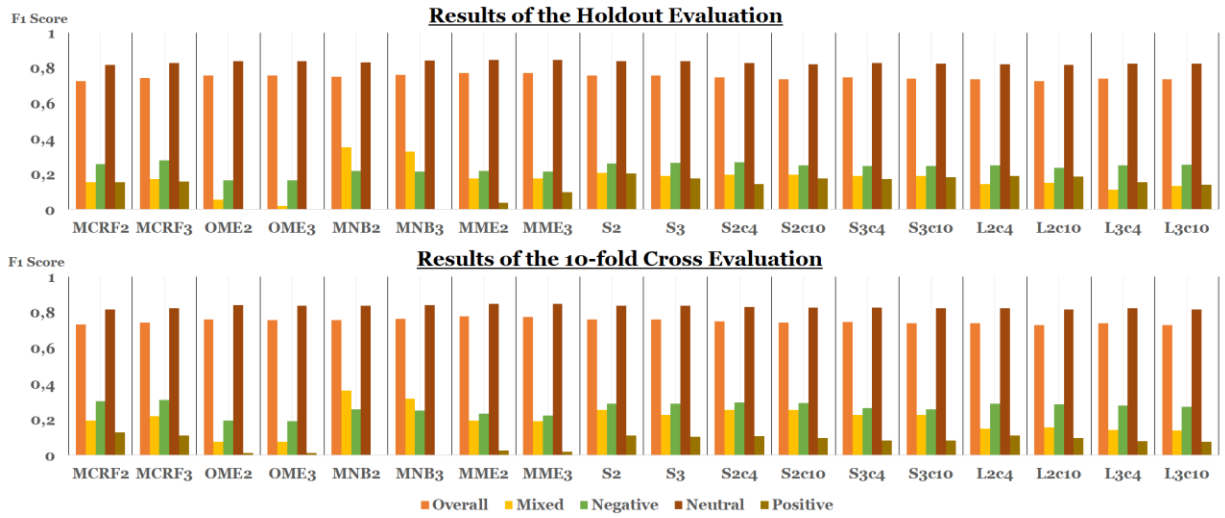

Figure 3. Results of the Evaluation

The pattern being that the neutral sentiment was largely the most accurately classified while positive sentiments were the least correctly classified. Further, mixed and negative sentiments exchanged positions sometimes depending on the algorithm. The results also show a large variance with the best F1 score being 0.85 and the worst being 0.13. Furthermore, some algorithms were shown to have a higher accuracy at classifying certain sentiments than others. Table 3 below contains the best performing algorithm-variation for each sentiment, the precision, recall and its F1 score.

Sentiment	10-fold Cross Evaluation				Holdout Evaluation			
	Algo.	Precision	Recall	F1	Algo.	Precision	Recall	F1
Positive	CRF_2	0.144	0.119	0.13	S_2	0.419	0.135	0.205
Mixed	NB_2	0.32	0.422	0.364	NB_2	0.305	0.423	0.354
Negative	CRF_3	0.37	0.265	0.309	CRF_3	0.355	0.228	0.278
Neutral	MME_2	0.822	0.88	0.85	MME_2	0.823	0.873	0.847
Overall	MME_2	0.812	0.748	0.778	MME_2	0.807	0.741	0.772

Table 3. Distribution of the Best Performing Algorithms

One can note that the overall best performing algorithm (MME_2) mimics that of the neutral sentiment in having higher F1 scores. This revealed a previously unknown skew of the dataset towards the neutral sentiment, due to the very large difference in class representation. A secondary analysis of the gold annotated data indicated that, out of the 19,863 annotated sentiments, approximately 82% were neutral sentiments, 13% negative, 3% mixed and 2% positive. With this background, the F1 score of about 0.2 for the positive sentiment looks to be a strong showing of the $LIBSVM_2$'s ability despite the stated amount of training instances for positive sentiments. A similar thing can be said of NB_2 with its F1 score of about 0.36

for the mixed sentiments. On the other hand, given the class representation of negative sentiments, its best algorithm's F1 score of about 0.3 indicates that negative sentiments were perhaps the most difficult to classify. Despite this, the consistent performance of MME_2 indicates an ability to generalize well and consistently given this size of the dataset. Apart from the negative sentiments, a closer look at the WNT revealed modifications did not make very significant difference. In most of the cases, a WNT of two was better in terms of performance by translating the input to a smaller feature set during training. Consequently, tuning it could help to reduce computational complexity and time taken for an algorithm to train a model, especially in terms of very large datasets.

Discussion

The results from our experiments show that it is useful to use the UIMA framework in creating an annotation pipeline for financial datasets. While no other contribution was found for a direct comparison, further literature exists, using even more complex NLP tools with ML algorithms for which the results of this approach compete adequately (Ferreira et al. 2014; O'Hare et al. 2009). However, in consideration of the obtained results and the particular car manufacturer use case, it turned out, that predictive analysis using neutral sentiments does not offer much potential. Having sentiments with a much clearer tendency, regardless whether they are positive or negative, could, for instance, drive the upward or downward trend of a stock index or better predict market response to a product. Due to this and the general lack of available human annotated or standardized datasets, further research would be very useful, especially in the financial domain and its investigations. The used demonstration case constitutes only one example, further evaluation are desirable at which for instance the ROC curve is examined in detail. Beyond that also other twerks, especially in case of the used algorithms are imaginable.

Conclusion

The financial and business industry requires in fast-moving and competitive times, like today, always novel ideas to identify the general sentiment of stakeholders. One promising and reliable source is represented by financial news, which are available in large unstructured quantities and of different trustworthiness. To overcome this situation, text mining methods appear to be a promising solution. Thus, in this work the current proliferation of different techniques, tools and procedures in providing annotations along with its challenges within the financial domain is investigated. By conducting the DSRM and different other methodology, the current state of the art is presented. In doing so, an artifact in the form of an annotation pipeline is presented, demonstrated and evaluated using UIMA, various machine learning algorithms and the case of five German car manufacturers. As results, this work provides beneficial insights for both researchers as well as practitioners, for instance to support decision making in terms of market response polarity analysis and stock polarity prediction.

REFERENCES

Albukhitan, S., Alnazer, A., and Helmy, T. 2018. "Semantic Annotation of Arabic Web Documents using Deep Learning," *Procedia computer science* (130), pp. 589–596.

Asooja, K., Bordea, G., and Buitelaar, P. 2016. "Using semantic frames for automatic annotation of regulatory texts," in *International Conference on Applications of Natural Language to Information Systems*, pp. 384–391.

Bethard, S., Ogren, P., and Becker, L. 2014. "ClearTK 2.0: Design patterns for machine learning in UIMA," in *International Conference on Language Resources & Evaluation*, p. 3289.

Bontcheva, K., and Cunningham, H. 2011. "Semantic annotations and retrieval: Manual, semiautomatic, and automatic generation," in *Handbook of semantic web technologies*: Springer, pp. 77–116.

Borg, M., Lennerstad, I., Ros, R., and Bjarnason, E. 2017. "On Using Active Learning and Self-Training when Mining Performance Discussions on Stack Overflow," in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, pp. 308–313.

Chakrabarti, A., Cormode, G., McGregor, A., and Thaler, J. 2014. "Annotations in data streams," *ACM Transactions on Algorithms (TALG)* (11:1), p. 7.

Chan, S. W. K., and Chong, M. W. C. 2017. "Sentiment analysis in financial texts," *Decision Support Systems* (94), pp. 53–64.

Chotipant, S., Hussain, F. K., Dong, H., and Hussain, O. K. 2015. "A neural network based approach for semantic service annotation," in *International Conference on Neural Information Processing*, pp. 292–300.

Cruz, N. P., Taboada, M., and Mitkov, R. 2016. "A machine-learning approach to negation and speculation detection for sentiment analysis," *Journal of the Association for Information Science and Technology* (67:9), pp. 2118–2136.
Cui, H., Boufford, D., and Selden, P. 2010. "Semantic annotation of biosystematics literature without training examples," *Journal of the American Society for Information Science and Technology* (61:3), pp. 522–542.
Das, A. S., Mehta, S., and Subramaniam, L. V. 2017. "AnnoFin-A hybrid algorithm to annotate financial text," *Expert Systems with Applications* (88), pp. 270–275.
Dasu, T., and Johnson, T. 2003. *Exploratory data mining and data cleaning:* John Wiley & Sons.
Davies, M. 2008. *The corpus of contemporary American English:* BYE, Brigham Young University.
Devitt, A., and Ahmad, K. 2007. "Sentiment polarity identification in financial news: A cohesion-based approach," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 984–991.
Diallo, G., Simonet, M., and Simonet, A. 2006. "An approach to automatic ontology-based annotation of biomedical texts," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 1024–1033.
Dickinson, M. 2015. "Detection of annotation errors in corpora," *Language and Linguistics Compass* (9:3), pp. 119–138.
Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and others 2003. "SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation," in *Proceedings of the 12th international conference on World Wide Web*, pp. 178–186.
Dingli, A. 2011. *Knowledge annotation: making implicit knowledge explicit:* Springer.
dos Santos, C., and Gatti, M. 2014. "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78.
El-Beltagy, S. R., Hazman, M., and Rafea, A. 2007. "Ontology based annotation of text segments," in *Proceedings of the 2007 ACM symposium on Applied computing*, pp. 1362–1367.
Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases," *AI magazine* (17:3), p. 37.
Ferreira, J. Z., Rodrigues, J., Cristo, M., and Oliveira, D. F. de 2014. "Multi-entity polarity analysis in financial documents," in *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, pp. 115–122.
Ferrucci, D., and Lally, A. 2004a. "Building an example application with the unstructured information management architecture," *IBM Systems Journal* (43:3), pp. 455–475.
Ferrucci, D., and Lally, A. 2004b. "UIMA: an architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering* (10:3-4), pp. 327–348.
Galke, L., Mai, F., Schelten, A., Brunsch, D., and Scherp, A. 2017. "Using Titles vs. Full-text as Source for Automated Semantic Document Annotation," in *Proceedings of the Knowledge Capture Conference*, p. 20.
Gao, G., Liu, Y.-S., Lin, P., Wang, M., Gu, M., and Yong, J.-H. 2017. "BIMTag: Concept-based automatic semantic annotation of online BIM product resources," *Advanced Engineering Informatics* (31), pp. 48–61.
Gotz, T., and Suhre, O. 2004. "Design and implementation of the UIMA Common Analysis System," *IBM Systems Journal* (43:3), pp. 476–489.
Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design science in information systems research," *MIS quarterly* (28:1), pp. 75–105.
Hotho, A., Nürnberger, A., and Paaß, G. 2005. "A brief survey of text mining," in *Ldv Forum*, pp. 19–62.
Hu, Z., Liu, W., Bian, J., Liu, X., and Liu, T.-Y. 2018. "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 261–269.
Ide, N., Fellbaum, C., Baker, C., and Passonneau, R. 2010. "The manually annotated sub-corpus: A community resource for and by the people," in *Proceedings of the ACL 2010 conference short papers*, pp. 68–73.
Kano, Y., Baumgartner Jr, W. A., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L., and Tsujii, J.'i. 2009. "U-Compare: share and compare text mining tools with UIMA," *Bioinformatics* (25:15), pp. 1997–1998.
Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., and Puppe, F. 2014. "UIMA Ruta workbench: rule-based text annotation," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pp. 29–33.
Krishnamoorthy, S. 2018. "Sentiment analysis of financial news articles using performance indicators," *Knowledge and Information Systems* (56:2), pp. 373–394.
Lesmo, L., Mazzei, A., and Radicioni, D. P. 2009. "Extracting semantic annotations from legal texts," in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pp. 167–172.
Liu, H., Wu, S., Tao, C., and Chute, C. 2012. "Modeling UIMA type system using web ontology language: towards interoperability among UIMA-based NLP tools," in *Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems*, pp. 31–36.
Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. 2011. "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pp. 142–150.
Mai, F., Galke, L., and Scherp, A. 2018. "Using Deep Learning For Title-Based Semantic Subject Indexing To Reach Competitive Performance to Full-Text," *arXiv preprint arXiv:1801.06717* .

Mayring, P. 2003. "Qualitative Inhaltsanalyse. Beltz," *Weinheim Basel* (8), pp. 5–135.
Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. 2015. "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," *Expert Systems with Applications* (42:1), pp. 306–324.
Ngai, E. W.T., and Lee, P. T.Y. 2016. "A Review of the literature on Applications of Text Mining in Policy Making," in *PACIS*, p. 343.
Njølstad, P. C. S., Høysæter, L. S., Wei, W., and Gulla, J. A. 2014. "Evaluating feature sets and classifiers for sentiment analysis of financial news," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, pp. 71–78.
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*, pp. 9–16.
Pan, F., Mulkar-Mehta, R., and Hobbs, J. R. 2011. "Annotating and learning event durations in text," *Computational Linguistics* (37:4), pp. 727–752.
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A design science research methodology for information systems research," *Journal of management information systems* (24:3), pp. 45–77.
Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P., and Alexander, M. 2017. "A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation," *Computer Speech & Language* (46), pp. 113–135.
Rahm, E., and Do, H. H. 2000. "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.* (23:4), pp. 3–13.
Ruiz-Martínez, María, Juana, Valencia-García, R., and García-Sánchez, F. 2012. "An Ontology-Based Opinion Mining Approach for the Financial Domain," in *Extended Semantic Web Conference*, pp. 73–86.
Sakurai, Y., Matsubara, Y., and Faloutsos, C. 2016. "Mining big time-series data on the Web," in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 1029–1032.
Salas-Zárate, M. D. P., Valencia-García, R., Ruiz-Martínez, A., and Colomo-Palacios, R. 2017. "Feature-based opinion mining in financial news: an ontology-driven approach," *Journal of Information Science* (43:4), pp. 458–479.
Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. 2012. "Evaluating sentiment in financial news articles," *Decision Support Systems* (53:3), pp. 458–464.
Shearer, C. 2000. "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing* (5:4).
Singh, G., Marshall, I. J., Thomas, J., Shawe-Taylor, J., and Wallace, B. C. 2017. "A Neural Candidate-Selector Architecture for Automatic Structured Clinical Text Annotation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1519–1528.
Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. 2013. "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.
Spina, D., Peetz, M.-H., and Rijke, M. de 2015. "Active learning for entity filtering in microblog streams," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 975–978.
Stilo, G., and Velardi, P. 2016. "Efficient temporal mining of micro-blog texts and its application to event discovery," *Data Mining and Knowledge Discovery* (30:2), pp. 372–402.
van de Kauter, M., Breesch, D., and Hoste, V. 2015a. "Fine-grained analysis of explicit and implicit sentiment in financial news articles," *Expert Systems with Applications* (42:11), pp. 4999–5010.
van de Kauter, M., Desmet, B., and Hoste, V. 2015b. "The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment," *Language resources and evaluation* (49:3), pp. 685–720.
Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., and others 2012. "A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools," *BMC bioinformatics* (13:1), p. 207.
Viani, N., Larizza, C., Tibollo, V., Napolitano, C., Priori, S. G., Bellazzi, R., and Sacchi, L. 2018. "Information extraction from Italian medical reports: An ontology-driven approach," *International journal of medical informatics* (111), pp. 140–148.
Wachsmuth, H., Rose, M., and Engels, G. 2013. "Automatic pipeline construction for real-time annotation," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 38–49.
Wang, J., Ren, F., and Li, L. 2014. "Recognizing sentiment of relations between entities in text," *IEEJ Transactions on Electrical and Electronic Engineering* (9:6), pp. 614–620.
Webster, J., and Watson, R. T. 2002. "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly*, pp. xiii–xxiii.
Wilcock, G. 2017. "The Evolution of Text Annotation Frameworks," in *Handbook of Linguistic Annotation*: Springer, pp. 193–207.
Zaidan, O., Eisner, J., and Piatko, C. 2007. "Using "annotator rationales" to improve machine learning for text categorization," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 260–267.