Sentiment Analysis

Introduction

In this day and age of modern computing, due to increase in use of social media, it has become increasingly difficult to analyze data. The Web is a large repository of data both structured and unstructured. First lets understand why do we need to do interpretation of data. Basically it can range from helping in knowing the perception of people regarding various topics to classifying numerous data points which consists of huge amount of data. This in turn helps to gather more knowledge about the pros and cons, which helps in improvement in the long run. Sentiment analysis has its applications in numerous fields. Where there is data involved there is always a sentiment. Opinion mining gives result which may be positive, negative or a neutral.

In this report, a detailed assessment of various works related to sentiment analysis is considered and explained. I studied in details, various approaches to perform a computational treatment of sentiments and opinions. We have analyzed the challenges and focused on understanding actual semantics of a particular document and numerous approaches that helps in achieving the efficiency of the classification.

Sentiment analysis otherwise known as Opinion Mining or text based mining refers to the analysis and extraction of semantic orientation. In the past few years, the rise of internet has been exponential which has led to increase in exchange of public opinion. And that is the main cause behind Sentiment Analysis today. And this does exploitation of opinions and sentiments of documents. The exploitation of this data to bring out public opinion and sentiment is very daunting task. The most fundamental problem here is sentiment polarity categorization. Sentiment analysis generally employs classification, but the document which has the contains the sentiment needs to be identified firstly.

Types of Sentiment Analysis:

There are basically four types:

1.Fine Grained: This type deals with the reviews which are in numeric form. For example ratings on a scale of 1 to 5 where 1 is very negative and 5 is very positive.

2.Aspect based: This is a improved version from the fine grained and hence gives us more information. It provides the specifics about what is lacking and

hence improvement can be done.
3. Emotion based: It deals with the emotion of the comment which can vary from happy to sad. Emotion is a integral part of the public review and this assessment hugely helps in applications like predicting genre of the movie used in recommender systems and music systems.
4. Intent based: This is the most important type of the opinion mining which has numerous use in the business field. This helps in analyzing the mood or intentions of the customers which eventually helps in the long run.

Use of Natural Language Processing:

Literature Survey

SENTIMENT ANALYSIS USING PRODUCT REVIEW DATA

In this paper, Xing Fang and Justin Zhan try to discuss the most fundamental problem in sentiment analysis, the sentiment polarity categorization, by considering a dataset containing over 5.1 million product reviews from Amazon.com with the products belonging to four categories: beauty, books, electronics and home. Although this paper tackles the problem of sentiment polarity categorization it still faces multiple challenges and has its limitations. One such being the curse of dimensionality in feature vector formation which limits the number of dimensions and also forces to have the same number of dimensions. The performance of this approach is estimated by considering the average F1 score. Therefore future work would be benefited if these limitations considered and thereby the accuracy and performance can be improved

EXPLORING PUBLIC PLACES FOR LIVEABLE PLACES BASED ON A CROWD-CALIBRATED SENTIMENT ANALYSIS MECHANISM

To fill the vacancy, a sentiment analysis service, called geo-sentiment analysis service is required. Thus, this paper firstly proposes CGSA: a Crowd-calibrated Geo-Sentiment Analysis mechanism, which can 1) start the sentiment analysis process based on the design of CTS (Compound Training Samples), and SSF (Social Sentiment Features), 2) perform three analyses, namely sentiment, clustering and time series analysis on geo-tagged social network messages, and 3) collect crowd-labelled data based on a crowdsourced calibration service to gradually improve the classification accuracy. SSF has the best accuracy in training sentiment classifiers, and the performance of the calibrated classifier increases gradually and significantly from 74.71namely 2D sentiment dashboard and 3D sentiment map, is implemented to support local authorities, urban designers and city planners better understand the effects of public sentiments regarding place (re)design in the test-bed area: Jurong East, Singapore. In general, three issues are solved, choosing sentiment features, maintaining up-to-date and localized lexicons. The application of social sentiment analysis brings many benefits in various domains, even using existing sentiment analysis tools, e.g. SentiStrength, or very simple analysis methods, which can only provide a very basic analysis without specific optimizations towards an application domain. Through this service, the service users, namely local authorities, urban designers and city planners, can better measure the satisfaction of people, and evaluate the fulfillment of predefined functionalities of facilities in the test-bed. Also, through these analyses, general sentiment patterns can be created, which can be used as baselines to evaluate the influence of changes or events, e.g., the reconstruction work, or the temporarily out of service of MRT (Mass Rapid Transit). Therefore, service users can better understand the reactions of people, and make better decisions.

EFFICIENT ADVERSE DRUG EVENT EXTRACTION USING TWITTER SENTIMENT ANALYSIS

In this paper, Yang Peng, Melody Moh, and Teng-Sheng Moh discuss how the advancements of social media are being helpful to extract large datasets by using a drug-related classification and sentiment analysis to extract ADEs on Twitter. ADEs are adverse drug events. Even though the pharmaceutical companies perform many drug-related tests beforehand, when a drug is released into the market some ADEs will be unidentified. Through the above-mentioned method, a data of four months on Twitter is collected so, as to capture the maximum number of ADEs. A simple and efficient pipeline is proposed to retrieve data from Twitter. The process of the pipeline is, the tweets from twitter are captured firstly and then the data is pre-processed (cleaned data is the output of data pre-processing). The drug classification is done for the cleaned data and the user opinion data is collected from which the ADEs are extracted. The captured tweets are stored in HIVE. Tweets are in JSON file and can, therefore, be stored in HIVE directly. They used python NLP tool for capturing tweets and Data pre-processing. For storing datasets of drug-related classification and tweets of sentiment analysis WEKA is used. Thus after thorough research on different tweets pipelines are built and they are compared to newly designed ones to extract numerous ADEs. As, a result an average of 5 times of total number of ADEs, among them 20

INVESTOR CLASSIFICATION AND SENTIMENT ANALYSIS

In this paper, Arijit Chatterjee and Dr. William Perrizo discuss the effect investors' bias has on the volatility of stocks in the market, sentiment analysis was done on tweets of the potential investors and also why they used Microsoft Azure over other sentiment analyser tools. Twitter is one of the largest social media platforms with over 280 million active users with almost 500 million tweets created every day. Some investors use Twitter to share their opinion on some ticker symbols every day, this paper discusses how these opinions of the investors affect the stock market. Investors are assumed to be sentiment driven. A top-down approach is used to make sure a stock is not overrated or underrated by the investors. The approach is based on two broad behavioral finance assumptions - sentiment and limits to arbitrage. Sentiment analysis is done on the tweets pulled from some selected investors' twitter feeds. They assign positive, negative and neutral sentiment scores to the ticker symbols from the pulled tweets by identifying "bad", "not good", "great" words in the tweets.

FORECASTING PRICE SHOCKS WITH SOCIAL ATTENTION AND SENTIMENT ANALYSIS

In this paper, the data from the Chinese Stock Market – SZSE and SSE are considered along with the social media activities in Weibo.com in order to extend recent studies on financial activities in social media and their impact on the stock market. What makes this work stand out is the way in which the previous limitations, such as, inability to tackle practical problems in finance, lack of proper knowledge regarding the direction of the price shock and, were overcome by using this implementation. This work makes use of DSA – Degree of Social Attention, which has been introduced by the previous works to capture stock price shocks. The method involves identifying the price shock as negative, near-zero or positive. These price shocks are essentially the difference between the expected value and actual value. Prior to estimating these price shocks, the social media activities are analysed and the features such as account information, future tracking, and response such as, like, repost, comment are considered. Furthermore, these details are cross-referenced to the account holder's actual activity and its effect on trading.

In [27], the researchers presented a subject sensitive sentiment analysis approach, which includes the context of tweets. According to authors the text cleansing techniques for input data before classification process can improve the results. Text cleansing includes normalization and vector representation of input data. They have pointed out that the subject aware classification brings the better results as compare to subject un-aware classification. The results can be further improved, if uni-gram approach is used instead of bi-gram or n-gram approach. A twitter dataset about word "Obama" was selected first. Features from tweets of selected dataset were extracted through Alchemy API, Tweet NLP and NTLK. From dataset, 30rest of 70for feature extraction and then the features were stored in a separate dictionary - Keyword$_Bundle -$
$inconjunctionwiththeirspecifictopicstoretainthetargetandcontextofthetweets.Thistechniquefurtherhel$

Conclusion

Used Apache PDFBox library to extract text and all additional information e.g. font type and font size from the financial prospectuses. Our constrained system uses only the provided training data without any additional external resources. Our system is based on the Maximum Entropy classifier and various features including font type and font size. Our system achieves F1 score 8177——— We approached the title detection subtask as a binary classification task. For all experiments, we use Maximum Entropy classifier with default settings from Brainy machine learning library. Maxent classifier is based on the Principle of

Maximum entropy.According to the principle of maximum entropy the distribution that represents the present state better has the maximum entropy. And on the basis of that the dataset gets choosen which has the maximum entropy. ———————— Povided training collection of documents contains the original documents in PDF format and an- notations JSON file with gold labels. The JSON file consists of an array of TOC items representing each title with the following properties: text - text of the title, id - order of occurrence the title, depth - depth level of the title, and page - page of title occurrence. After extraction use our own algorithm to link the annotations to the extracted text representation. ach line of text a separate text segment and classify each segment as title or non-title. If there is a change in the font size or type we split the text into two lines. Additional metadata are extracted from the first occurring word of the given line. The metadata include the following features: Is bold, Is italic, Is all caps, Begins with cap, Begins with numbering, Left position, Font size, and Font type ———————————— following features proved useful and were used in our submissions. • Character n-grams (ChNn): Separate feature for each n-gram representing the n-gram presence in the text. We do it separately for different orders n 1, 2 and remove n-gram with frequency f 2. • Binary Features (B): We use separate binary feature for the following text characteris- tics (Is bold, Is italic, Is all caps, Begins with cap, Begins with numbering, Is next line empty, Is prev line empty). • Position Features (P): We use four separate binary features to represent the difference in the left position of the text for two sentences. The positions can be equal, lower, greater, and missing. We compare sentence at position p with sentence at position p 2, p 1, and p + 1. • First Orto-characters (FO): Bag of first three orthographic1 characters with at least 2 occurrences. • Last Orto-characters (LO): Bag of last three orthographic1 characters with at least 2 occurrences. • Font Size (FS): We map the font size of text into a one-hot vector with length twelve and use this vector as features for the classifier. The frequency belongs to one of twelve equal-frequency bins2. Each bin corresponds to a position in the vector. We remove font sizes with frequency 2. • Font Size Diff (FSD): We use four separate binary features to represent the difference in font size (FS) of the text for two sentences. The positions can be equal, lower, greater, and missing. We compare sentence at position p with sentence at position p 1 and p + 1. • Font Type Diff (FTD): We use three separate binary features to represent the difference in font type for two sentences. The font type can be equal, different, and missing. We compare sentence at position p with sentence at position p 1 and p + 1. • Font Type Unigrams (FTU): We tokenize font type name and use the presence of unigrams as a feature we remove unigrams with frequency f 1000. ——————- resutls: performed ablation experiments to illustrate which features are the most beneficial (see Table 3). Numbers represent the performance change when the given feature is removed (i.e. lower number means better feature). We used approximately 30used the rest of the dataset for training the features We also repeated the experiment using leave-one-out cross-validation as the previous experiment seemed inaccurate. Our evaluation measure is macro-averaged F1-score. We can see that the experiments are inconclusive as some of the findings are in contradiction. The most helpful features in terms of leave-one-out cross-validation apart from character bi-grams include both first and last orto-characters and font type unigrams. Last orto-characters, FSD and FTD were always beneficial. On the contrary position features and font size features were the least helpful features.

Future work

They presented a technique that is based on the fact that the given a large corpus, geometric layout and character features of the titles and non-titles can be learnt separately and their learning can then be transferred to a domain-specific dataset. On a domain-specific dataset, along with a pre-trained model, they train a Deep Neural Net on the text of the document for geometric and character features. This approach achieved an F-Score of 83.25. F score: calculated as the harmonic mean of the above two terms precision and recall. Precision:This is the ratio of True positive to sum of True positive and False Positive. This signifies how relevant the results which have been obtained. recall: This is the ratio of True positive to the sum of True positive and False negative. This signifies how rightly classification is done. ——————————————— Be it title detection or TOc,Understating the inherent document layout and structure benefits several downstream document AI tasks such as search, summarizing, entity extraction and table detection Humans glance at a document and comprehend the document structure including the titles vs non-titles as well as the overall hierarchy of the titles. Humans have intuitive notions of how a document is structured and the assumption is confirmed af- ter reading a text block. which the machine lacks So,Transfer learning can be used to model the structural properties of a general document. They took Arxiv documents1 available in the open-domain to learn the structural model of a general document.

——————————————— Literature on title detection can be classified broadly into three categories: 1.works that deal with ToC page of documents, works that use images of document pages and works which use the geometrical and textual features of the text blocks.

1. ToC pages of documents, after the ToC pages are detected, the title entries are extracted and mapped to the pages by finding links between title and corresponding pages. El Haj et al.(2014) used this approach in detecting titles in UK Financial Reports. As they rely on ToC pages, they cannot be applied to documents that do not have ToC pages.

2.so approaches use computer vision to fragment the page image into entities such as text, title and table. used Convolutional neural networks combined with graphical models to identify the entities in a document page

3.some approaches use learning or rule-based methods to detect headers based on textual and geometrical features. usually used in digitally generated documents like webpages and native PDF documents. ——————————————— pre-process the PDF files by converting them to XML documents by Poppler2 library. These files are then parsed to merge elements similar in styling and located in close proximity. Headers and Footers are identified and removed by page association (Lin, 2003) as they would hinder the process of title detection. Our proposed title detection method has three components; Pre-trained neural network to model gen- eral structural information, Sequential Network to learn domain-specific text and training of both the networks combined. The network comprises 22 manual features as depicted in Table 1. Model is trained by multi-layer neural network as described by the architecture ———- Dataset and Training We take around 6000 Arxiv documents from the annotated documents provided by Muhammad Mah- bubur Rahman and Tim Finin (2017). The data split is shown in Table 4. The training was done for three models, namely, geometric, character and character plus geometric. Character plus geometric model performed the best as expected. Intuition being features from geometric and character will complement each other when trained together. We got a significant rise of 5Training metrics are depicted in Table ————————————————————- We use LSTM as a sequence classification model. Intuition being common phrases that are part of financial titles can be learnt by a sequence network such as LSTM. We use Glove word vector embedding. Last word cell state is passed as input to two dense layers. Final layer after the dense layer performs title detection Table 5. Out Best F Score on the test dataset was at 73. 3.2.1 Dataset and Training The architecture is mentioned in Table 5. Dataset Split can be seen in Table 4. Our best performing model on validation set gave 73bidirectional and attention mechanisms (Abi Akl et al., 2019) . ———————————————————— Full network comprises of pre-trained weights from Character plus Geometric Model trained on Arxiv PDF and Sequence Model trained on FinToC dataset. Last dense layer from Sequence Model and Char- acter Plus Geometric Model are concatenated. One more and last dense Layer of 10 units is added after that. Loss function is binary cross entropy. Total trainable parameters are 2, 26, 335. No layers is freezed for subsequent training. 3.3.1 Dataset and Training FinToC Dataset as mentioned in Table 4 was used. Adam optimizer, epochs equal to 30 and batch size of 500 were used as hyperparameters in the model. The code was written in Tensorflow v1.15 (Abadi et al., 2015) Jointly trained final architecture got the F-Score of 83.25 on the test set. ———————————————————— Results: Two highlights of the final model are • Pre-trained weights captured the generic structure of documents, giving a boost to accuracy. This transfer learning approach can be improved further by using better architectures and features which are domain-independent.This procedure achieved a 10• Combination of geometric and character-based features complemented each other to attain higher accuracy compared to either of them separately.

They submitted two systems for the final evaluation. • First one is the Joint Trained LSTM and Geometric+CharCNN network Second one was the ensemble of the first one and an XGBoost(Chen and Guestrin, 2016) model

References

[1].Fang, Xing Zhan, Justin. (2015). Sentiment analysis using product review data. J Big Data. 2. 10.1186/s40537-015-0015-2.
[2]. L. You and B. Tunçer, ”Exploring public sentiments for livable places based on a crowd-calibrated sentiment analysis mechanism,” 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016, pp. 693-700, doi: 10.1109/ASONAM.2016.7752312.
[3].Moh, M., Moh, TS., Peng, Y. et al. On adverse drug event extractions using twitter sentiment analysis. Netw Model Anal Health Inform Bioinforma 6, 18 (2017). https://doi.org/10.1007/s13721-017-0159-4
[4].Chatterjee, Arijit Perrizo, William. (2016). Investor classification and sentiment analysis. 1177-1180. 10.1109/ASONAM.2016.7752388. [5].Keli Xiao, Qi Liu, Chuanren Liu, and Hui Xiong. 2017. Price Shock Detection With an Influence-Based Model of Social Attention. ¡i¿ACM Trans. Manage. Inf. Syst.¡/i¿ 9, 1, Article 2 (February 2018), 21 pages. DOI:https://doi.org/10.1145/3131781