NLG301 at SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News

Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{cjchen, hhhuang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

Short length, multi-targets, target relationship, monetary expressions, and outside reference are characteristics of financial tweets. This paper proposes methods to extract target spans from a tweet and its referencing web page. Total 15 publicly available sentiment dictionaries and one sentiment dictionary constructed from training set, containing sentiment scores in binary or real numbers, are used to compute the sentiment scores of text spans. Moreover, the correlation coefficients of the price return between any two stocks are learned with the price data from Bloomberg. They are used to capture the relationships between the interesting target and other stocks mentioned in a tweet. The best result of our method in both subtask are 56.68% and 55.43%, evaluated by evaluation method 2.

1. Introduction

Nowadays the discussion of finance on social media such as twitter reveals the feeling of a market in some degrees. Market sentiment analysis becomes an important financial technology in trading strategies (Kazemian, 2014). Financial tweet sentiment classification differs from traditional sentiment classification in several ways. Firstly, the sentiment degree is a real number rather than discrete numbers such as 1 (positive), 0 (neutral), and -1 (negative).

Secondly, a financial tweet usually concerns multiple targets. In the tweet, “Oil To Break Out: Adding Chevron https://t.co/IrZkAVxjiE $AXP $CLGRF $CSCO $ERX $IBM $MCD $SSRI $VLO $WMT $XOM $CVX”, all companies denoted by the cashtag $ticker-symbol share only one description. The activity, oil to break out, is a good news for energy companies, but may be bad for shipping companies. Domain knowledge about the targets is necessary for suitable interpretation in this case.

Thirdly, numbers in the financial tweets are quite important. In the tweet, “MarketWatch: RT wmwitkowski: Guess who sold off about $800 million in $MDLZ after losing about $1 billion on $VRX???https://t.co/SHiJutyenv”, large number means more negative. In contrast, in the tweet (named T1 hereafter), “$AAPL now up 2.2% w/div since my original call, while $SPY up only 0.6% even w/ this Fri's div. #EMH be damned. Still holding”, the larger the number is, the more positive score the target will get. The activity related to numbers determines the polarity and its degree.

Fourthly, sentiment scores depend on the activity of the companies, and their relationships, e.g., the adversarial relation versus the cooperate relation. In the tweet, “Report: Apple signs up for Google's cloud, uses much less of Amazon's $AAPL $GOOG $GOOGL $AMZN $DROPB https://t.co/zN3KDGYvGT”, $AAPL $GOOGL $AMZN and $DROPB are assigned sentiment scores 0.15, 0.443, -0.38 and -0.213, respectively, by human annotators because Amazon and Dropbox are two competitors of Google in the cloud market.

This paper explores various types of features selected from the text span related to the interesting targets for fine-grained financial tweet sentiment classification. Both human and machine labelled text spans are used and compared. This paper is organized as follows. Section 2 surveys the related work. Section 3 presents the identification of text spans and extraction of features from them.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 847–851,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

Section 4 shows and discusses the experimental results. Section 5 concludes the remarks.

2. Related Work

Go et al. (2009) employ Naïve Bayes, Maximum Entropy, and SVM to classify sentiment of Twitter messages to positive, neutral, and negative categories. Usernames, usage of links and repeated letters are taken as features. Jiang et al. (2011) consider target-dependent features and related tweets in the target-dependent Twitter sentiment classification, and achieve an accuracy of 68.2%. The sentiments of the tweets are still discrete, i.e., positive, negative, or neutral.

Takala et al. (2014) develop an evaluation dataset for topic-specific sentiment analysis in financial and economic domain, where financial news are sampled from Thomson Reuters newswire. Each news story is annotated by 7-point scale from very positive to very negative. The SemEval-2017 Task 5 deals with fine-grained sentiment analysis on financial microblogs and news. Financial tweets and news headlines are taken as evaluation data.

This paper is different from the above coarse-grained approaches. Multi-targets in a short text is one of the major issues to be tackled. We will find the sentiment of an interesting target in a tweet

3. Features

The twitter dataset in SemEval-2017 Task 5 is used in this study. It consists of 1,539 financial tweets. Total 55.6% of tweets contain more than one target. The sentiment scores of the targets in a tweet are labeled into the real numbers between -1 to 1 by 3 experts.

Figure 1 shows the structure of a financial tweet. The following sections will discuss how to extract features from each component. The 21 features used in the experiments are shown in Table 1.

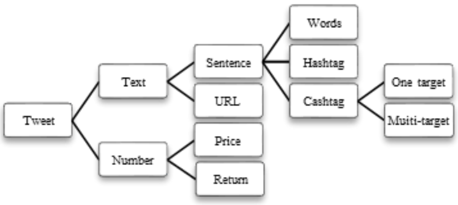

Figure 1. Structure of a financial tweet

Type	Source	Dictionary/Operations	#Features
text span	in tweet	15 dictionaries/average	15
text span	in tweet	trained dict/avg or max	2
text span	via URL	SenticNet 4/avg or max	2
number	in tweet	not applicable	1
relation	in tweet	not applicable	1

Table 1. Features

3.1 Text Span

In the SemEval-2017 dataset, experts labeled a "text span" for each target. In the tweet T1 specified in Section 1, the spans "$SPY up only 0.6%, Still holding" and "now up 2.2%, Still holding" are assigned to $SPY and $AAPL, respectively.

Besides human-annotated text span, we also explore two span determination algorithms shown as follows. Here, we collect a mapping table of tickers and company names from Bloomberg, and use it to find the canonical form of the company mentions in a tweet.

(1) Position-based approach (EP)

First, a tweet will be separated into sentences. Then, the sentence containing the cashtag or the company name of a target is deemed as the span for the target. The sentence without any company names or cashtags will be regarded as the span for all targets shown in the tweet. Here, a hashtag is regarded as a word.

(2) Dependency-based approach

(2.1) Stanford Parser (ED-S)

A tweet is parsed by Stanford dependency parser (Marneffe, 2006). To reduce the effects of out of vocabulary (OOV) words in parsing, cashtags and company names are replaced by common names like "Bob". A dependency tree for n-word tweet is composed of n triples in the form of dep(word$_i$,word$_j$), where word$_i$ and word$_j$ has a dependency dep, word$_i$ is a parent of word$_j$, and word$_j$ is a child of word$_i$. We take the ancestors and the decedents of a target as its span.

(2.2) TweeboParser (ED-T)

TweeboParser is a dependency parser, designed for tweets (Kong et al., 2014). It tries to deal with the following challenges: token selection, multi-word expressions, multiple roots, and structure within noun phrases. The multiple roots property tends to provide shorter span than ED-S with the same extracted algorithm.

The average length of the tweets, manual spans, EP spans, ED-S and ED-T spans are 17.61, 6.26, 12.17, 10.27, and 7.78 words, respectively. Compared with 140-character word limit in twitter, a financial tweet is very short. In particular, manual spans are much shorter.

Besides text span in tweets, tweets may contain URLs as reference (shown in Figure 1). To collect as much information as possible, we parse the web page designated by the URL and retrieve the sentences containing the target. Those sentences are considered as additional text for sentiment classification.

Table 2 shows the example of span for the single target case: “$ATVI ooks pretty bullish for now. from a short-term perspective, it's got a good chance of maybe sliding back to 33.70 #stocks #investing”

Table 3 shows the example of span for: “Report: Apple signs up for Google's cloud, uses much less of Amazon's $AAPL $GOOG $GOOGL $AMZN $DROPB https://t.co/zN3KDGYvGT” The target of span is $AAPL.

As shown in the above two examples, ED-T method provides the shortest span and extracts the span more similar to Manual span. The extract results may sometimes include all words in tweet.

Manual	ooks pretty bullish for now
EP	'$ATVI ooks pretty bullish for now.', 'from a short-term perspective', " it's got a good chance of maybe sliding back to 33.70 #stocks #investing"
ED-S	Bob ooks pretty bullish for now from a short term perspective it s got a good chance of maybe sliding back to 33 70 stocks investing
ED-T	$atvi ooks pretty bullish for now

Table 2. Example of Single Target Span

Manual	Report: Apple signs up for Google's cloud
EP	Report: Apple signs up for Google's cloud, uses much less of Amazon's
ED-S	Report Bob signs up for Google s cloud uses much less
ED-T	apple signs up for google's cloud

Table 3. Example of Multi-Target Span

3.2 Ensemble of Sentiment Dictionaries

In the lexicon-based sentiment analysis, the sentiment score of a text span is determined by the sentiment scores of the sentiment words it contains. We use the max or the average of the sentiment scores of the related words as the features shown in Table 1. Total 15 sentiment dictionaries of two forms, real value and binary, are consulted. In addition to the publicly available dictionaries, we also construct a sentiment dictionary from the training set automatically.

(1) Real value: SentiWordNet[1], SenticNet 4[2], NRC Hashtag Emotion Lexicon, NRC Hashtag Affirmative Context Sentiment Lexicon and NRC Hashtag Negated Context Sentiment Lexicon unigrams and bigrams, Yelp Restaurant Sentiment Lexicon unigrams and bigrams, Amazon Laptop Sentiment Lexicon unigrams and bigrams, Sentiment140 Affirmative Context Lexicon unigrams and bigrams, Emoticon Lexicon aka Sentiment140 Lexicon unigrams, bigrams[3]

(2) Binary (1 for positive, -1 for negative): NRC Word-Emotion Association Lexicon, Macquarie Semantic Orientation Lexicon

SentiWordNet is quite different from the other sentiment dictionaries in the above. Words of different senses are assigned different sentiment scores. In the experiments, we use Babelfy (Moro et al., 2014) to disambiguate the word senses before consulting SentiWordNet.

We separated all words in training set, then counted the average sentiment score for each word to construct the other sentiment dictionary.

[1] Download from http://sentiwordnet.isti.cnr.it/

[2] Download from http://sentic.net/

[3] Other dictionaries can download from http://saifmohammad.com/WebPages/lexicons.html

3.3 Numbers and Relationships

As described in Section 1, numbers such as monetary expressions are important cues for analyzing financial tweets. We use the position between a target and numbers to decide which number in a tweet is related to a target, and calculate the ratio of this number and sum of all numbers mentioned in the tweet as a feature.

In finance, the correlation coefficient of the price return between two stocks has been used to capture their relationship. We calculate the correlation coefficients by the stock prices during 2015/01/01 to 2016/10/31 downloaded from Bloomberg, and use them to compute the geometric mean of the absolute value of the correlation coefficients between the interesting target and other stocks in a tweet. The sign of this feature depends on the plurality voting.

4. Experimental Results

In coarse-grained classification, we classify the sentiment of a given target mentioned in a tweet into positive, negative, or neutral. The SVM model with the proposed 21 features are used. In fine-grained classification, we predict the sentiment score of the given target in real number. The SVR is adopt for this task. Both model followed the default parameters used in python Scikit-learn (Pedregosa et al., 2011) Accuracy and cosine similarity are used to measure the performance of the coarse-grained and the fine-grained tasks, respectively. The ground truth is represented as a vector of targets' sentiment scores in cosine measurement. Four-fold cross-validation is conducted.

Table 4 shows the experimental results. Using manual spans achieves the ideal performance because the critical text span for the target is known beforehand. Using EP spans is better than using the complete tweet, but worse than using manual spans. The performance of the ED-S Span approach does not meet our original expectation due to the noise results in dependency parsing in tweets, which are usually incomplete sentences. The ensemble of the first five methods show in Table 4 show the best accuracy in the 4-fold validation for both coarse-grained and fine-grained case.

Methods	Coarse-grained	Fine-grained
Complete Tweet	64.91%	59.40%
Manual Span	86.94%	82.17%
EP Span	72.00%	62.83%
ED-S Span	64.52%	39.48%
ED-T Span	34.50%	34.05%
Ensemble	89.60%	82.60%

Table 4. Experimental Results

Although ED-T method provides the span whose average length is the closest to manual span, it gets the worse accuracy. It's worth to leave no stone unturned. We leave this part in the future works.

Due to the limited amount of submission, we submit two test results: manual span and the ensemble. The final result for the SemEval-2017 Task 5 by cosine similarity are 35.66% and 38.28%, and by evaluation method 2[4] are 55.34% and 56.68%, as the test of 4-fold validation that ensemble result got the best accuracy.

The same dictionaries and SVR model are used to subtask 2, using the news headline data. The best result is 55.43% using evaluation method 2.

5. Conclusion and Future Work

In this paper, we analyze the specific properties of the financial tweets, and propose methods to extract features from a tweet and its mentioned URL.

We illustrate some of the challenges of analyzing financial tweets. For the multi-targets problem, in order to extract the specific part of tweet for the target, we provide three methods, making the process automatically. The comparison of "spans" similarity will be provided in the future works. Moreover, we will also handle the number and the relationship between targets more precisely.

[4] Description of evaluation method 2 : http://alt.qcri.org/semeval2017/task5/data/uploads/description_second_approach.pdf

References

Alec Go, Richa Bhayani and Lei Huang. 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1, 12

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 151–160

Siavash Kazemian, Shunan Zhao, and Gerald Penn.2014. Evaluating sentiment analysis evaluation: A case study in securities trading. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 119–127

Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. Proceedings of Ninth International Conference on Language Resources and Evaluation, pages 2152–2157

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A Dependency Parser for Tweets. In EMNLP, 1001–1012

A. Moro, A. Raganato, R. Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics (TACL), 2, pp. 231-244.

Pedregosa et al, 2011. Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), pages 2200–2204, Valletta.

E. Cambria et al., "SenticNet 4: A Semantic Resource for Sentiment Analysis based on Conceptual Primitives," Proc. 26th Int'l Conf. Computational Linguistics, 2016, pp. 2666–2677

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 29 (3), 436-465

Svetlana Kiritchenko, Xiaodan Zhu and Saif Mohammad. 2014. Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research, volume 50, pages 723-762

Kiritchenko, S., Zhu, X., Cherry, C., 2014 Mohammad, S.M.: NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval '14, pp. 437–442

Svetlana Kiritchenko, Xiaodan Zhu and Saif Mohammad.2014. Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research, volume 50, pages 723-762

Saif Mohammad and Peter Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon, In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010

Saif Mohammad, Bonnie Dorr, and Cody Dunne. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)

Cortis, Keith and Freitas, Andr and Dauert, Tobias and Huerlimann, Manuela and Zarrouk, Manel and Handschuh, Siegfried and Davis, Brian. 2017. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).