Classifying facts and opinions in Twitter messages: a deep learning-based approach

Swayambhu Chatterjee, Shuyuan Deng, Jun Liu, Ronghua Shan & Wu Jiao

Published online: 23 Aug 2018.

Submit your article to this journal

Article views: 4829

View related articles

View Crossmark data

Citing articles: 6 View citing articles

THE OPERATIONAL RESEARCH SOCIETY

Taylor & Francis
Taylor & Francis Group

Check for updates

Classifying facts and opinions in Twitter messages: a deep learning-based approach

Swayambhu Chatterjee[a], Shuyuan Deng[b], Jun Liu[a], Ronghua Shan[a] and Wu Jiao[c]

[a]College of Business and Information Systems, Dakota State University, Madison, SD, USA; [b]Seidman College of Business, Grand Valley State University, Grand Rapids, MI, USA; [c]College of Business Administration, Northern Illinois University, DeKalb, IL, USA

ABSTRACT

Massive social media data present businesses with an immense opportunity to extract useful insights. However, social media messages typically consist of both facts and opinions, posing a challenge to analytics applications that focus more on either facts and opinions. Distinguishing facts and opinionss may significantly improve subsequent analytics tasks. In this study, we propose a deep learning-based algorithm that automatically separates facts from opinions in Twitter messages. The algorithm outperformed multiple popular baselines in an experiment we conducted. We further applied the proposed algorithm to track customer complaints and found that it indeed benefits subsequent analytics applications.

1. Introduction

The past decade has seen an explosive growth of user-generated content through social media platforms. As of 2015, leading social media websites such as Facebook has already been generating 250 million posts per hour and Twitter has been generating more than 21 million messages per hour (Simos, 2015). This staggering data volume presents an immense opportunity for businesses to understand their customers and competitors through analytics (He, Wu, Yan, Akula, & Shen, 2015; Mikalef, Pappas, Krogstie, & Giannakos, 2017). For many organisations, opinion extracted from user-generated content has become an invaluable asset that helps determine the success of a product in the market (Wright, 2009). Opinion mining, the field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language, has been playing a key role in the analytics of the vast amount of social media data (Liu, 2012). For instance, user opinion on social media has been used to predict stock returns (Bollen, Mao, & Zeng, 2011; Luo, Zhang, & Duan, 2013), music sales (Dewan & Ramaprasad, 2014), and election results (Tumasjan, Sprenger, Sandner, & Welpe, 2011). On the other hand, real-time social media message streams can also be used for the early detection of events. For example, Li, Lei, Khadiwala, and Chang (2012) designed a Twitter-based event detection and analysis system (TEDAS) to uncover important events and also retrieve related tweets. Paul and Dredze (2011) plotted the number of disease-related topics

on Twitter against the actual reported level of diseases and found a close match between the two series.

Social media analytics, using computational approaches to derive insights from the massive volumes of online user-generated data, has become a primary application of big data analytics today. Despite the wide adoption of social media analytics, one task that has been largely overlooked in the literature is to distinguish between opinions and facts among social media messages. According to Liu (2010), facts are objective expressions about entities, events, and their properties, whereas opinions are usually subjective expressions that describe people's sentiments, appraisals, or feelings towards entities, events, and their properties. Consistently, in the context of social media, we consider a *fact* as a user narrative that reports an event objectively and an *opinion* as a statement manifesting a user's personal belief or sentiment. In the current landscape of social media blogosphere, especially in microblogs like Twitter, messages related to an event or entity of interest usually consists a mixed corpus of opinions and facts. Analytics applications that focus on either extracting facts or opinions are facing an enormous amount of noisy data. Distinguishing between opinions and facts in user-generated content also offers advantages for many natural language processing applications, guiding decisions about what information to extract and how to organise it (Yu & Hatzivassiloglou, 2003). Certain consumers of social media data are more interested in learning facts in real time while others focus on user opinions. For example, during the 2014–2016 Ebola outbreak in West Africa, a large amount of Twitter messages contains the

---

word, *Ebola*. Health agencies who are monitoring the spread of infectious diseases may pay special attention to newly reported cases of "Ebola" among these messages. For this particular task, users' sentiment due to the outbreak is less relevant to their objectives. However, other consumers of social media data such as airlines and tourism agencies may be more interested in assessing the public fear towards an Ebola outbreak since their operations will be affected by such opinions.

To address the inevitable challenge of having to deal with both facts and opinions in social media analytics, this study demonstrates a new algorithm that can separate facts and opinions in Twitter messages. Specifically, we propose a set of manually engineered features for the supervised classification of facts and opinions. We then combined these features and the bag-of-words (BOW) features using a deep neural network (DNN). We benchmark its performance using a labelled dataset and compare it with popular subjectivity detection and sentiment analysis tools. We further illustrate the usefulness of the proposed algorithm in a real-world business scenario and show that the signal from social media is enhanced after separating opinions from facts. To ensure rigour and relevance, we adopted the design science methodology throughout the paper (Hevner, March, Park, & Ram, 2004; March & Smith, 1995; Nunamaker Jr, Chen, & Purdin, 1990).

This study makes several contributions to social media analytics research and practice by: (1) demonstrating the relevance of distinguishing facts and opinions, (2) developing a novel deep learning algorithm integrating two types of features to classify facts and opinions in Twitter messages, (3) evaluating the algorithm by comparing it with existing subjectivity detection methods and other popular baselines, and (4) evaluating the usefulness of the algorithm in a real-world business scenario. The findings in this study have important implications on how companies can better utilise massive volumes of social media data in terms of analytics to support their decision-making which may potentially help them gain a competitive advantage.

The rest of the paper is organised as follows: Section 2 reviews related studies on social media analytics, prior attempts to distinguish facts and opinions, and their limitations. Section 3 examines the available information in Twitter messages and provides the rationale for the manually engineered features that are useful for classifying facts and opinions. It then describes the deep learning architecture we propose in this study. Section 4 evaluates the proposed design and compares its performance with existing popular solutions. Section 5 further illustrates the relevance of the proposed design in a practical opinion mining application that tracks emerging customer complaints from Twitter messages. Section 6 concludes the study by discussing the implications of the findings to research and practice, the limitations, and future directions.

2. Related work

There has been a wide range of applications developed over the last decade that mine social media data for diverse objectives to obtain meaningful insights. In terms of data consumption, we can broadly classify this active research area into two categories. The first category, event detection, represents research that captures factual data, such as identifying breaking news from social media streams. The second category, opinion mining, aims at capturing user opinion and/or towards events or entities. In this section, we review some of the important and representative studies in both directions. We then review the existing subjectivity detection and text classification methods and identify their limitations in classifying facts and opinions.

2.1. Event detection

Facts discussed in online communities largely reflect real-world events. Detecting occurrences of real-world events from social media is a very active and relevant research domain. It usually takes hours or days for a traditional media to report an event, depending on the importance of the event (Zhao et al., 2013). Many events that are not significant enough to the general public but of great interest to special groups may not get covered at all. Many prior studies focused on detecting events by analysing Twitter messages. For instance, Li et al. (2012) designed a TEDAS to uncover important events from tweets. Becker, Iter, Naaman, and Gravano (2012) and (Ritter, Etzioni, & Clark, 2012) examined how to retrieve relevant social media messages related to a known event and proposed methods to identify and summarise important events or popular topics from tweet streams. Marcus et al. (2011) designed a system, TwitInfo, to summarise and visualise information on Twitter. Factual narrative posted by users in social media is widely used by news media. It is already a fixture for reporting for many journalists, especially around breaking news events where non-professionals may already be on the scene to share an eyewitness report, photo, or video of the event (Diakopoulos, 2012). Overall, this stream of research focuses on identifying and summarising news on social media. Despite its usefulness, it does not attempt to distinguish opinions and facts in the data related to an event.

2.2. Opinion mining

Discovering opinion from social media messages is another active research area in social media analytics. Companies use sentiment analysis to develop marketing strategies by assessing and predicting

public attitudes towards their brand (Cambria, Schuller, Xia, & Havasi, 2013). Some of the earlier seminal studies include evaluation of review sentiment (Turney, 2002), product reputation (Nasukawa & Yi, 2003), and assessing investment opportunity (Das & Chen, 2007). With the growth in social media data, such as Facebook and Twitter messages, capturing and analysing public opinion has become more pervasive. Examples include predicting commercial success of music (Dewan & Ramaprasad, 2014) and predicting firm value based on blog metrics (Luo et al., 2013). The past decade has also seen significant advances in sentiment analysis tools for social media data, such as SentiStrength (Thelwall, Buckley, & Paltoglou, 2012) and VADER (Hutto & Gilbert, 2014). Despite their effectiveness in classifying sentiment orientation, they do not attempt to distinguish facts and opinions.

2.3. Subjectivity detection

Subjectivity detection is a task that identifies subjective contents from textual data. It can be considered as an early attempt to separate facts and opinions. Research on subjectivity detection has been active for a couple of decades in the field of information systems, computer science, journalism, sociology, and political science. Earlier work (e.g., E. Hatzivassiloglou & Wiebe, 2000; Riloff & Wiebe, 2003; Wiebe, 2002) used BOW and related features, as well as lexicon-based methods in detecting subjective statements in news articles and extended reviews. Lexicon-based methods compare each word in a document to a predefined set word list (Ding, Liu, & Yu, 2008; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). One of the most popular tools for subjectivity detection is OpinionFinder, which relies on a lexicon containing a list of common subjective words (Wilson et al., 2005). When we try to apply subjectivity detection methods on social media data, especially microblog messages, two challenges emerge. First, microblog messages are short and informal. Emoticons, abbreviations, and hashtags are frequently used to convey an important part of the meaning. Subjectivity detection techniques that heavily depend on textual information may not provide satisfactory results when applied to social media texts. Second, facts can also be represented in a subjective manner. Similarity, not all subjective statements are opinions. Consider the following tweet,

Ebola Economics: As Liberia, Sierra Leone Bounce Back, Women And Youth Still Struggle – http://t.co/uS0XhijpKz *#ebola*

It seems to express a subjective belief due to the presence of the terms, *bounce back* and *struggle*. However, it in fact includes a news title and the link to the news article. The article describes a series of facts and statistics related to the Ebola outbreak in certain regions of Africa. Thus, what this tweet reports should be considered as a fact. Take another tweet for example, *villages are abandoned in Sierra Leone*. It seems to present a strong negative opinion according to the widely used MPQA lexicon due to the presence of the word *abandoned*. However, this sentence reports a fact without expressing much user opinion. Thus, the current subjectivity detection approaches are often ineffective when a fact contains subjective cues.

2.4. Text classification

Text classification aims at categorising documents into predefined categories (Chou, Sinha, & Zhao, 2010; Feldman & Sanger, 2007; He et al., 2015). Popular applications of text classification in the IS field include fraud detection (e.g., Abbasi, Zhang, Zimbra, Chen, & Nunamaker Jr, 2010; Siering, Koch, & Deokar, 2016; Zhou, Burgoon, Twitchell, Qin, & Nunamaker Jr, 2004) and sentiment classification (Oh & Sheng, 2011; Yu, Duan, & Cao, 2013; Ghiassi, Zimbra, & Lee, 2016). To extract features from texts, the most common technique is the BOW model, representing each document as a vector of word frequencies. However, BOW models, when used on short texts such Twitter messages, suffer from the feature sparsity problem, where each document may only contain a few non-zero fields among a large feature space (Deng, Sinha, & Zhao, 2017b). Social media messages are also typically informal. The various types of abbreviations and misspellings make the feature space even larger. Thus, comparing to longer documents such as news articles and online reviews, a short text like a Twitter message represented using BOW may not provide sufficient information for a classification task. Lexicon-based methods also have limitations when applied to Twitter data due to the prevalent acronyms, abbreviations, slang words, and misspelled words in Twitter message (Deng, Sinha, & Zhao, 2017a). For example, the tweet, *ebola is scarrryy*, clearly represents negative sentiment. However, with only a lexicon-based method, the token *scarrryy* is not likely to match any entry in the predetermined lexicon, consequently creating a low-recall problem.

2.5. Research gaps

In sum, the existing applications of event detection and opinion mining applications have largely overlooked the task of distinguishing facts and opinions in social media messages. The existing subjectivity detection and popular text classification methods have obvious limitations when directly applied to this task. These research gaps have motivated us to propose a novel algorithm for classifying facts and opinions in social media data. We are particularly

interested in analysing Twitter messages since they have been extremely popular in social media analytics. Focusing on Twitter data allows us to maximise the impact of our findings and their business values in a variety of domains.

3. Classifying facts and opinions in twitter messages

In this section, we describe the design of a deep learning architecture that effectively combines BOW features and our hand-engineered features for classifying facts and opinions in Twitter messages. BOW features are essential in text classification (Joulin, Grave, Bojanowski, & Mikolov, 2016). Supplementing text information with manually engineered features may overcome the limitations of BOW features in classifying short texts. Thus, we use both types of features for classifying facts and opinions in Twitter messages.

A normal approach for combining feature set would be to concatenate them before using a machine learning model. However, a BOW representation may contain thousands of features even for a small corpus. The number of manually engineered features is usually quite limited. Directly combining them may put too much weight on the BOW features thus diminish the effect of the manually engineered features. To address this issue, we propose a DNN that accepts each type of features separately and combines them in a hidden layer. In the following subsections, we describe the manually engineered features first and then present the deep learning model.

3.1. Twitter metadata

The challenge of short-text classification has driven us to explore features external to the text. Twitter provides a rich set of metadata or the contextual information about the posted message. The metadata contain information ranging from the user's profile to the geographical location when the message was posted. Such information can be broadly categorised into four components, the tweet, the user, entities, and places. To overcome the noisy nature of social media data, some studies (e.g., Castillo, Mendoza, & Poblete, 2011; Diakopoulos, 2012; Sankaranarayanan, Samet, Teitler, Lieberman, & Sperling, 2009; Wang, Wang, Li, Abrahams, & Fan, 2014) have successfully demonstrated the effectiveness of metadata for a diverse set of objectives. In our study, we also leverage Twitter metadata for classifying facts and opinions.

3.2. Feature engineering

As aforementioned, classification of short-text messages is a challenging task if we only use textual information. Thus, we utilise on features based on both the text and the metadata for classifying facts and opinions. In machine learning, features are often either automatically learned and manually engineered based on expert knowledge or heuristics. In this subsection, we describe the features we manually engineered based on heuristics from analysing a developing corpus that is independent of our test bed. All of these features are binary-coded.

3.2.1. Feature 1: title capitalisation

Twitter messages that aim at conveying factual information are usually written more formally. They are usually free of spelling errors. Such messages concisely summarise the fact owing to the character limitation of the platform. A key indicator of such formal and summarised information is the title capitalisation, in which the first letter of each important word is capitalised. Consider the following tweet, *UN: Ebola Still Global Emergency Despite Big Drop in Cases – ABC News.. Related Articles*: *http://t.co/zD21Tl3Fw3*. The message essentially consists of the title of the news article and the link to it. Thus, we consider this feature useful in distinguishing between factual and opinionated content. We calculated the number of words with only the first letter in uppercase. If more than three words satisfy the criterion, we assign a 1 for this binary feature.

3.2.2. Feature 2: URL

We have observed that in social media, especially in Twitter, factual narratives are often accompanied by an external reference. Consider the following tweet, *Bank of America Merrill Lynch out volunteering at Roadrunner Food Bank this morning. https://t.co/QhfzKUAS0Z*. The user reports an actual event with an external link to substantiate the claim. Thus, we create a binary variable indicating if a Twitter message contains an URL.

3.2.3. Feature 3: user type

Narratives in social media that originates from institutional users like news networks and corporations are usually factual in nature. The tweet attribute, *user description*, provides some useful information in this regard. For example, a user description of *Link news about Africa* is highly likely to be interested to capture and disseminate news or factual content about Africa. Thus, we check the content of the user description field using a list of news agency names. To create this list, we collected a reference dataset of more than 300 publicly available media accounts (shown in Appendix A).

3.2.4. Feature 4: followers

Factual content posted on Twitter are usually targeting at a generic and diverse audience. Such accounts usually have a larger number of followers. For example, the official Twitter account of CNN has over 37

million followers. Thus, we created this variable by checking if the number of followers of the user is greater than 500.

3.2.5. *Feature 5: mention of news agency*

We also observed from the developing corpus that tweets mentioning a news agency's account tend to provide a factual content or at the least aims at substantiating the narrative with a verifiable source of information. Sometimes, they just retweet a message posted by a news agency's account, such as ..*RT @japantimes: Heads of Ebola-hit nations meet Obama*. This type of posts, although not providing a complete URL, propagates factual or reported content. Thus, we check if any news agency is mentioned in tweet text based on the list created for Feature 3.

3.2.6. *Feature 6: numbers*

We also observed that posts containing numbers often tend to be factual in nature. Consider the post ..*World Bank must lead efforts to raise $1.7 billion to improve Ebola-hit countries health care http://t.co/D4xSz0k1uv #humanrights*. Presence of numbers is also observed in posts containing date time or currency information. Thus, we check if a tweet contains numbers.

3.2.7. *Feature 7: repeating characters*

The presence of repeating characters usually emphasises an opinion or sentiment. Consider the following tweet, *Ebola is scarrryy*. The repetition of the letter *r* emphasises the strong sentiment or opinion about the entity ebola by the user. To create this variable, check if any character is repeated more than twice in a tweet.

3.2.8. *Feature 8: all uppercase*

Like repeating characters or multiple special characters, capitalising all letters in a word usually indicate an opinion. To create this feature, we check if a tweet contains words with all letters capitalised.

3.2.9. *Feature 9: twe0072minology*

We also observed that presence of Twerminology (Internet slang used on Twitter) usually indicates an opinion due to its informality. To create this variable, we used an external reference dataset (Appendix B) comprising more than 100 Twitter slang words to check the presence of such words in the Twitter messages.

3.3. *The deep learning model*

Deep learning models are typically used to learn a multi-level representation of data (LeCun, Bengio, & Hinton, 2015). Such deep architectures, often consisting multiple hidden layers, can learn more complex relationships between variables (Evermann, Rehse, & Fettke, 2017; Kraus & Feuerriegel, 2017). The hidden layers are learned from data rather than specified by human experts. Such hierarchy of hidden layers can often capture discriminative information and suppress irrelevant variations. For instance, multi-layer perceptron is an early application of deep learning. Given the advances of graphics processing unit (GPU) computing and optimisation algorithms in the past decade, more complex deep learning frameworks have been made possible and lead to significant improvement in image recognition and natural language processing (Schmidhuber, 2015).

To effectively combine the hand-engineered features and the BOW features, we utilise a deep learning model with two branches to combine their representation in a hidden layer. The left branch of the model utilises a fully connected (dense) layer to learn the representation of the manually engineered features. The layer consists of 100 nodes and is followed by a dropout layer. In the dense layer, each input is mapped to each of the 100 nodes through a rectified linear unit (ReLU). ReLU uses the higher between the input and 0 as the output and is the most frequently used activation function among deep learning models (LeCun et al., 2015). The dense layer is then followed by a dropout layer. Dropout refers to removing nodes and its connections from a neural network (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The use of dropout can reduce both overfitting and computation complexity. The choice of dropout nodes is random. In this study, we use a dropout probability of 0.5 since it is believed to be close to optimal for a variety of applications (Srivastava et al., 2014). The right branch of the model learns the representation of the BOW features. It uses a dense layer of 200 nodes and is also followed by a dropout layer with a dropout probability of 0.5. The outputs from both branches are then combined using a merge layer through horizontal concatenation. We then use a one-node dense layer with sigmoid activation to output the classification result. Figure 1 illustrates this architecture using a diagram. Each layer is represented using a rectangle with the layer type, number of nodes, and the activation function (when applicable).

4. Evaluation – classifying fact and opinion

We evaluate the usefulness of the proposed algorithm in two phases. This section evaluates the algorithm's performance classifying facts and opinions in Twitter messages. In the next section, we evaluate its business value using a real-world business scenario.

4.1. *Data*

Using Twitters Application Programming Interface, we collected Twitter messages using the keyword "Ebola" for a short period of time. As discussed in the Introduction, health agencies may be interested in monitoring new cases of Ebola (fact) through Twitter
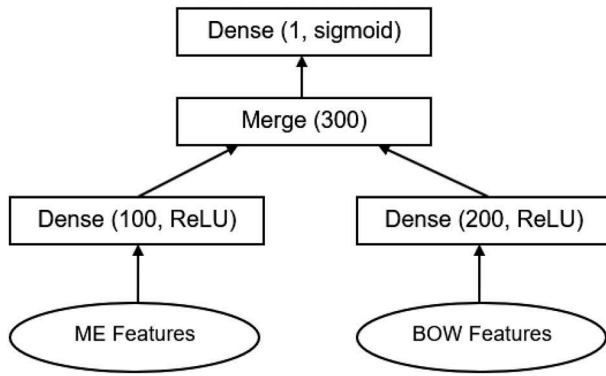
Figure 1. The proposed DNN.

messages while travel agencies and airlines may be interested in assessing the public fear towards Ebola (opinion). Thus, this search term is relevant for our study. As a part of preprocessing, we filtered out all non-English tweets since the language resources (e.g., Twitter slang word list) we use are specifically for the English language. We then eliminated duplicate texts. We also filtered out tweets that less than three words or only contain an URL. As a result, we obtained 2268 unique tweets.

4.2. Data labelling and reliability

To evaluate the classification correctness, we manually created a labelled dataset. We initially invited two independent raters to manually label each tweet as a fact or an opinion. The basic definition of fact and opinion were given to the raters. A Twitter message may contain both opinion and facts. In this case, we consider it as an opinion. The raters are graduate students in the Information Systems major and are very familiar with Twitter.

Inter-rater reliability (IRR) provides a way of quantifying the degree of agreement between two or more raters who make independent ratings on a set of subjects. We computed Cohen's Kappa to estimate IRR (Cohen, 1960). The Kappa statistic is 0.71, indicating substantial reliability (Landis & Koch, 1977). Consistent with prior research (Aggarwal, Gopal, Gupta, & Singh, 2012), when the two raters disagreed on the label of a message, a third rater's judgment was used. In our labelled dataset, 937 of the tweets were classified as facts and 1331 were classified as opinions.

4.3. Baselines

We selected several baseline models to compare against the proposed method. The first is OpinionFinder (Ellen Riloff, Wiebe, & Wilson, 2003; Wiebe & Riloff, 2005), an application that identifies subjective sentences and sentiment expressions from documents. It was one of the earliest attempts to detect subjective sentences. For each tweet detected with a subjective sentence, we consider the tweet as an opinion; otherwise, fact. The second baseline is SentiStrength (Thelwall et al., 2012), a superior sentiment analysis tool for Twitter messages (Abbasi, Hassan, & Dhar, 2014). It generates a positive score and a negative score for each tweet. We computed a net score and consider tweets with a net score of zero as fact and otherwise opinion.

We also use popular text classifiers and BOW features, the manually engineered features (ME), and their combination as baselines. The four classifiers used are linear support vector machine (SVM), logistic regression (LR), extreme gradient boosting (XGB), and random forest (RF). SVM has long been data scientists' favourite in learning high-dimension data given its robustness in preventing overfitting. LR is also used because it can be considered a neural network without the hidden layers, which is directly comparable to our proposed algorithm. XGB (Chen & Guestrin, 2016) is an ensemble classifier based on tree boosting which has won numerous data mining competitions in recent years. RF is another superior ensemble classifier that we think might be useful in our classification task.

4.4. Results

We performed 10-fold cross validation for each of the algorithms using the labelled tweets. The proposed DNN was run using the Keras[1] (Tensorflow[2] backend) package in Python on a desktop computer with a GTX 1080 TI GPU. Table 1 reports the accuracy, precision, recall, and F-measure (in percentage) of classifying facts and opinions using our labelled data set. OpinionFinder and SentiStrength did not perform very well with both F-measures below 50%. Among the classifiers using only BOW features, the highest accuracy (85.23%) and F-measure (82.52%) were achieved by LR. Among the classifiers using only the ME features, the highest accuracy (73.85%) and F-measure (70.97%) were achieved by XGB. Comparing these two sets of results, it appears that BOW features are still essential in our classification task.

For the group of classifiers using the combination of ME and BOW features, our proposed DNN performed best in terms of accuracy (87.08%) and F-measure (85.20). The SVM, LR, XGB, and RF classifiers using the combined features do not show much improvement compared to those using only BOW features. This suggests that our proposed algorithm that combines the representation of different types of features works significantly better than concatenating the feature at the input stage. Overall, our proposed algorithm is most effective in classifying facts and

Table 1. Results of classifying facts and opinions in Twitter messages (%)

Features	Classifier	Accuracy	Precision	Recall	F-Measure
ME + BOW	DNN	87.08	81.73	89.33	85.20
	SVM	84.61	78.91	86.54	82.29
	LR	84.97	79.19	87.19	82.80
	XGB	83.07	74.41	90.50	81.58
	RF	82.01	77.67	80.89	78.75
ME	SVM	73.10	66.41	71.07	68.54
	LR	73.32	67.15	69.89	68.40
	XGB	73.85	65.72	77.36	70.97
	RF	72.62	64.26	76.40	69.69
BOW	SVM	84.35	80.35	83.03	81.43
	LR	85.23	81.27	84.41	82.52
	XGB	84.09	77.98	86.34	81.68
	RF	84.57	83.34	78.85	80.80
Standalone	OpinionFinder	46.29	62.29	48.52	37.78
	SentiStrength	50.63	41.69	48.77	44.95

opinions in Twitter messages compared with a variety of baselines.

5. An application – assessing regulatory risk

In this section, we illustrate how distinguishing facts and opinions can be useful in practical business applications. In particular, we show that using only opinion tweets, rather than a mixed stream of both facts and opinion, can be used to better identify emerging consumer complaints, mitigating regulatory risk for financial institutions.

5.1. Background

Preventing consumer dissatisfaction from escalating into official complaints is an important task for many businesses. These customer complaints often result in punitive actions by federal authorities and may cost millions of dollars in fines. For instance, after the 2008 financial crisis, federal regulators have been closely monitoring every practice area of financial institutions. As of 2015, banks have been charged with more than $320 billion dollars for behaviour deemed questionable by federal authorities (Finch, 2017). Consequently, organisations are spending a significant amount of resources to identify potential regulatory risk. It is imperative for financial institutions to understand their customers' concerns before they result in any punitive actions by the federal authorities. Apart from facing penalties from regulators, listening and understanding customer demands is an important task for any business to improve their product and service offerings as well as to maintain their reputation among competitors.

The Consumer Financial Protection Bureau (CFPB) supervises financial institutions to ensure their compliance with federal consumer financial laws, to assess risks to consumers, and to help ensure a fair and transparent marketplace for consumers (CFPB, 2016). The Dodd-Frank Wall Street Reform and Consumer Protection Act, whose passage in 2010 was a legislative response to the financial crisis of 2007–2008 and the subsequent Great Recession (Eaglesham, 2011), lead to CFPB's creation. Since 2011, CFPB opened their public web interface to capture consumer grievances online. On receiving formal complaints from customers, the bureau seeks out a response from the concerned company about their specific practice area and on getting an unsatisfactory response. It may also initiate a legal proceeding against that company, which often results in punitive action and settlement of huge amount.

User opinions reflected in Twitter messages can be used to assess the level of customer dissatisfaction which may later result in a formal complaint to CFPB. In general, when there are more negative opinions, the risk is likely to be higher. Thus, this application may be more effective if opinions are separated from facts on Twitter. Consider the following two tweets, "Bank of America continues its 5-day losing streak currently down 1.43% ($BAC)" and "@jesseltaylor STUPID Bank of America reopened my wife's CLOSED account instead of just letting a $20 auto-debit bounce.". Both messages are negative about Bank of America. However, the former represents a negative fact regarding the loss of its stock value. It does not indicate any potential CFPB complaint. On the other hand, the latter expresses a consumer's dissatisfaction with Bank of America's customer service. It could potentially turn into a formal complaint about the bank's account management practice.

5.2. Method and result

Prior studies have associated user opinions on social media with a variety of social and economic outcomes, such as stock returns (e.g., Bollen et al., 2011), movie sales (e.g., Mishne & Glance, 2006), and music sales (e.g., Dewan & Ramaprasad, 2014). In the same vein, we designed a process that correlates Twitter sentiment with CFPB complaints. First, we collected Twitter messages mentioning a financial institution. Second, we classified the sentiment of each tweet and count the number of negative messages each week. Third, we calculated the Pearson correlation between the weekly negative tweets and weekly CFPB complaints for the company. Fourth, we applied the proposed algorithm to classify each tweet as fact or opinion and calculated the correlation between the negative opinion tweets and the CFPB complaints. By comparing the correlations from Step Three and Step Four, we can see if using only opinion tweets classified by the proposed algorithm is more effective in early monitoring consumer complaints than using both fact and opinion tweets.

We collected Twitter messages mentioning a major financial institution, Banks of America, for 16 weeks.

We used SentiStrength to classify these tweets into positive, negative, and neutral. We then classified the negative tweets as facts and opinions using the proposed algorithm. The weekly correlation between all negative tweets and the actual CFPB complaints was 0.50. When using only negative opinion tweets, the correlation increased to 0.54. This result suggests that distinguishing facts and opinion using the proposed algorithm can indeed benefit business applications.

6. Discussion and conclusion

In this study, we proposed a deep learning algorithm that combines BOW features and manually engineered features for classifying facts and opinions in Twitter messages. We further demonstrated that distinguishing facts and opinions using these features can indeed benefit practical opinion mining applications. This study makes several contributions to analytics research and practice. First, we identified the importance of distinguishing facts and opinions in social media analytics. Social media platforms are not only a pool of user opinions but also a popular news channel. However, existing opinion mining methods made little attempts to separate opinion from facts when extracting user opinion from social media data. Our findings show that distinguishing between fact and opinion can indeed improve related social media analytics applications which increases its value for businesses. Second, drawing on the recent advances in deep learning, we proposed a DNN that effectively combines manually engineered features and BOW features in a hidden layer. The classification results show that combining the representation of features can indeed outperform combining the features directly. This finding has important implications for a wide range of analytics applications. Third, we elucidated the practical relevance of the proposed algorithm using a real-world opinion mining scenario, assessing the risk of CFPB complaints using Twitter sentiment. Financial institutions can directly benefit from this finding by adopting our algorithm in their social media analytics practice. In the same vein, other social media analytics applications, such as stock prediction and event detection, can also benefit from the same procedure.

This study is not without limitations. First, the opinion mining application that correlates negative opinion tweets with CFPB complaints is more a proof-of-concept than a complete solution. Future research can enhance such an application by incorporating predictive models. They can also test its robustness using more companies over a longer period. When testing the proposed features in classifying facts and opinions, we used a sample of 2268 labelled Twitter messages. This has limited the number of layers in our neural network model. Future studies can use large datasets to take advantage of supervised deep learning models by incorporating more hidden layers.

Notes

1. https://keras.io.
2. https://www.tensorflow.org.

Acknowledgements

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abbasi, A., Hassan, A., & Dhar, M. 2014. *Benchmarking Twitter sentiment analysis tools. Proceedings of the ninth international conference on language resources and evaluation*,Reykjavik, Iceland. Paris, France: European Languages Resources Association (ELRA).

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker Jr, J. F. (2010). Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*, 34(3), 435–461.

Aggarwal, R., Gopal, R., Gupta, A., & Singh, H. (2012). Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth. *Information Systems Research*, 23(3–Part–2), 976–992.

Becker, H., Iter, D., Naaman, M., & Gravano, L. 2012. *Identifying content for planned events across social media sites. Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 533–542). New York, NY: ACM.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 2, 15–21.

Castillo, C., Mendoza, M., & Poblete, B. 2011. *Information credibility on twitter. Proceedings of the 20th international conference on World wide web*(pp. 675–684). New York, NY: ACM.

CFPB. (2016). Consumer Financial Protection Bureau: Enforcing federal consumer protection laws. Retrieved from http://files.consumerfinance.gov/f/documents/07132016_cfpb_SEFL_anniversary_factsheet.pdf

Chen, T., & Guestrin, C. 2016. *Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining*(pp. 785–794). New York, NY: ACM.

Chou, C.-H., Sinha, A. P., & Zhao, H. (2010). A hybrid attribute selection approach for text classification. *Journal of the Association for Information Systems*, 11 (9), 491–518.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. [Article]. *Management Science*, 53(9), 1375–1388.

Deng, S., Sinha, A. P., & Zhao, H. (2017a). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, 94, 65–76.

Deng, S., Sinha, A. P., & Zhao, H. (2017b). Resolving ambiguity in sentiment classification: The role of dependency features. *ACM Transactions Manage Information Systems*, 8(2–3), 4–13.

Dewan, S., & Ramaprasad, J. (2014). Social media, traditional media, and music sales. [Article]. *MIS Quarterly*, 38(1), 101–121.

Diakopoulos, N. C. M. D. N. M. 2012. *Finding and assessing social media information sources in the context of journalism. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2451–2460). New York, NY: ACM.

Ding, X., Liu, B., & Yu, P. S. 2008. *A holistic lexicon-based approach to opinion mining. Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240). New York, NY: ACM.

Eaglesham, J. (2011). Warning shot on financial protection. Retrieved from https://www.wsj.com/articles/SB10001424052748703507804576130370862263258?mod=googlenews_wsj.

Evermann, J., Rehse, J.-R., & Fettke, P. (2017). Predicting process behaviour using deep learning. *Decision Support Systems*, 100, 129–140.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, UK: Cambridge University Press.

Finch, G. (2017). World's biggest banks fined $321 billion since financial crisis. Retrieved from https://www.bloomberg.com/news/articles/2017-03-02/world-s-biggest-banks-fined-321-billion-since-financial-crisis

Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems*, 33(4), 1034–1058.

Hatzivassiloglou, V., & Wiebe, J. M. 2000. *Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of the 18th conference on computational linguistics* (Vol. 1, pp. 299–305). Stroudsburg, PA: Association for Computational Linguistics.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7), 801–812.

Hevner, V. A., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.

Hutto, C. J., & Gilbert, E. E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. Paper presented at the eighth international conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI,

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv Preprint arXiv*, 1607, 01759.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104, 38–48.

Landis, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Li, R., Lei, K. H., Khadiwala, R., & Chang, K.-C.-C. 2012. *Tedas: A twitter-based event detection and analysis system. Data engineering (ICDE), 2012 IEEE 28th international conference on* (pp. 1273–1276). Washington, DC: IEEE.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2, 627–666.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167.

Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research*, 24(1), 146–163.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251–266.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. 2011. *Twitinfo: Aggregating and visualizing microblogs for event exploration. Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 227–236). New York, NY: ACM.

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2017). Big data analytics capabilities: A systematic literature review and research agenda. *Information Systems and e-Business Management*, 1–32. doi:10.1007/s10257-017-0362-y

Mishne, G., & Glance, N. 2006. *Predicting movie sales from blogger sentiment. Proceedings of AAAI 2006 Spring symposium on computational approaches to analysing weblogs* (pp. 155-158). Menlo Park, CA: AAAI Press.

Nasukawa, T., & Yi, J. 2003. *Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on Knowledge capture* (pp. 70–77). New York, NY: ACM.

Nunamaker Jr, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. *Journal of Management Information Systems*, 7(3), 89–106.

Oh, C., & Sheng, O. R. L. 2011. *Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. The International conference on information systems* (pp. 57–58). Atlanta, GA: Association for Information Systems.

Paul, M. J., & Dredze, M. 2011. *You are what you tweet: Analyzing twitter for public health. Fifth International AAAI conference on weblogs and social media (ICWSM 2011)*. Menlo Park, CA: AAAI Press.

Riloff, E., & Wiebe, J. 2003. *Learning extraction patterns for subjective expressions. Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 105–112). Stroudsburg, PA: Association for Computational Linguistics.

Riloff, E., Wiebe, J., & Wilson, T. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003-Volume 4*, 2003 (pp. 25–32): Stroudsburg, PA: Association for Computational Linguistics

Ritter, A., Etzioni, O., & Clark, S. 2012. *Open domain event extraction from twitter. Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1104–1112). New York, NY: ACM.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. 2009. *Twitterstand: News in tweets.*

Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 42–51) New York, NY: ACM.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117.

Siering, M., Koch, J.-A., & Deokar, A. V. (2016). Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems, 33* (2), 421–455.

Simos, G. (2015). How much data is generated every minute on social media? Retrieved from http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/2015

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology, 63*(1), 163–173.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review, 29*(4), 402–418.

Turney, P. D. 2002.*Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA: Association for Computational Linguistics.

Wang, G. A., Wang, H. J., Li, J., Abrahams, A. S., & Fan, W. (2014). An analytical framework for understanding knowledge-sharing processes in online Q&A

communities. *ACM Transactions Manage Information Systems, 5*(4), 1–31.

Wiebe, J. (2002). Instructions for annotating opinions in newspaper articles.

Wiebe, J., & Riloff, E. 2005. *Creating subjective and objective sentence classifiers from unannotated texts. International conference on intelligent text processing and computational linguistics* (pp. 486–497). Berlin/Heidelberg, Germany: Springer.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. 2005. *OpinionFinder: A system for subjectivity analysis. Proceedings of HLT/EMNLP on interactive demonstrations* (pp. 34–35). Stroudsburg, PA: Association for Computational Linguistics.

Wright, A. (2009). Our sentiments, exactly. *Communications of the ACM, 52*(4), 14–15.

Yu, H., & Hatzivassiloglou, V. 2003. *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 129–136). Stroudsburg, PA: Association for Computational Linguistics.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919–926.

Zhao, L., Chen, F., Dai, J., Hua, T., Lu, C., & Ramakrishnan, N. (2013). Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS One, 9*(10), e110206–e110206.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker Jr, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139–166.

Appendix A. Twitter account names of news agencies

ABCNews	Ars Technica	big_picture
abc7	arstechnica	Billboard.com
ABC7 News	BBC Breaking News	billboarddotcom
ABCNews.com	BBC Click	Bloomberg
ABCNewsLive	BBC World News	Boing Boing
AJELive	bbcbreaking	boingboing
AJEnglish	bbcclick	BRACworld
AL Jazeera	BBCNews	Breaking News
AL Jazeera English	bbcworld	breakingnews
CBS News	gizmodo	NY Times
CBS TOP News	Guardian Tech	nytimesworld
cbsnews	guardiannews	nytimesphoto
CBSTopNews	guardiantech	nytpolitics
China News	GdnDevelopment	parisreview
CNET News.com	Harvard Business Rev	Paste Magazine
cnetnews	harvardbiz	pastemagazine
cnn	Huffington Post	People magazine
CNN Breaking News	huffingtonpost	peoplemag
CNN Live	Inc. Magazine	Pete Cashmore
cnnbrk	incmagazine	Pitchfork
cnnlive	instyle	pitchforkmedia
cnnmoney	io9	planetmoney
CNNMoney.com	japantimes	politico
cracked	latimes	Radiolab
Cracked.com	nytimesworld	reuters
CW Network	LIFE.com	Richard MacManus
cw_network	lifehacker	Rolling Stone
Daily Telegraph News	Los Angeles Times	rollingstone
dailynation	mashable	rww
davos	McKinsey Quarterly	TNN
DISCOVER Magazine	mckquarterly	sciam
discovermag	Mental Floss	Science Channel
Drudge Report	mental_floss	Science News
drudge_report	msnbc	sciencechannel
E! News	msnbc_breaking	sciencenewsorg
E! Online	mtv	Scientific American
emarketer	natgeosociety	SeattleTimes
eMarketer	National Geographic	Seed Magazine
Empire Magazine	New Scientist	seedmag
empiremagazine	New York Post	Sky News
enews	New York Times Photo	SkyNewsBreak
engadget	newscientist	Slashdot
Engadget	newsweek	slate
eonline	newyorkpost	Smashing Magazine
espn	Nielsen Wire	smashingmag
Fast Company	nielsenwire	Social Media Today
fastcompany	NPR News	socialmedia2day
finabiz	NPR Politics	Talking Points Memo
Fox News	NPR's Planet Money	techcrunch
FOX10Phoenix	nprnews	telegraphnews
foxnews	nprpolitics	The Big Picture
freakonomics	NY Review of Books	The Economist
freakonomics	nybooks	guardian
gawker	New York Times	The Guardian
gdgt	nytimes	The Moment
The New York Times	US News	finabiz
The Onion	Us Weekly	Nollywood news
The Paris Review	USA Breaking News	IBNMoney
The Today Show	USA TODAY Top News	BIGLAWNewsLine
The Washington Post	usabreakingnews	BreitbartNews
The Weather Channel	usatoday	SkyNewsBreak
The White House	usweekly	Nigerian News
theeconomist	VANITY FAIR	bignewsnetwork
themoment	vanityfairmag	france24
theonion	waitwait	Fox News
TIME.com	Wall Street Journal	International News
todayshow	washingtonpost	NYCnowApp
tpmmedia	Washington Times	DunyaNews
Travel + Leisure	whitehouse	CBS Local
travlandleisure	wired	newsofthesouth
TV Guide	wnycradiolab	Sierra Leone times
tvguide	World Economic Forum	drudgereport
tvnewser	wsj	Ghana news
TVNewser	Breaking News	Eyewitness
TWCBreaking	WSJbreakingnews	news
tweetmeme	yahoo	tribune
UNICEF	YahooNews	

Appendix B. Twitter slangs

#FF	FOE	LAGO	OH	TGTSIO!
AAP	FTW	LMFO	OTB	TMB
BM	HBL	LOT	PPP	TQRT
CBOT	HWEVR	LYS	REDLYFE	TWIT
DNA	IOT	MOV	RTHX	TYFF
F4F	JGH	MVO	SML	WBOS
FMOT	LA	NEWEETER	SWEEPLE	WMBY
FTC	LLBLOG	OOYF	TCAT	WTFS
HAB	LOML	POLITWEE	TGIC	XTB
HT	LU	QOTP	TJP	YOCO
IMF	MIWI	RT	TPL	BTW
JATA	MUVA	SA	TWARS	CHK
KYR	NEJ	SMHD	TWTR	BFN
LGY	OOTT	SWAG	WATN	OMG
LOL	PMOT	TC	WMBU	OMG
LTOTD	PWAT	TGFAD	WTF	ZZZZZ
MISTWEET	RQB	TINGLE	WWTT	FCUK
MULC	S2G	TON	YOBK	DA
NEF	SHXP	TW	YTS	FAB
OOMF	STOW	TWITTERE	AYEG	F2F
PART	TBT	VMA	BTW	ETA
PTP	TEOG	WKYP	DM	EMA
RLRT	TIL	WTB	ELW	IDK
S/S	TNJ	WWD	FGF	IC
SFTW	TTYLXOX	YNI	FTAD	KK
STBH	TWITT	YOLTBTST	GSTG	NTS
T1Z	UOT	775	HR	TWART
TCOT	WGD	AWS	IGERS	
THOT	WPR	BTTP	IWTKCR	
TMC	WWBA	DETWEET	KOT	
TSS	YIF	EHT	LFL	
TWITE	YOLO	FDU	LOAL	
UAYA	1V4	FSTOW	LRT	
WDILY	AGIG	GLWS	MFBTY	
WOMW	BTTE	HMU	MT	
WTWT	CMSU	IDTS	NCBD	
YBM	EHE	ITGFT	OLLI	
YOFO	FBTL	KFB	PAP	
#LESM	FSF	LBOD	PRT	
AF	GBP	LMR	RIH	
BOL	HDM	LPC	RTZ	
CCW	IDGHP	MBF	SERPT	
DWEET	IT	MRT	SPEET	
FATC	JK	MYWB	TCIC	