3_Sentiment_Analysis

Nicholas Mitchell

March 30, 2016

Contents

1 Sentiment Analysis

1.1 Sentiment analysis: definition and origins

Also termed *opinion mining*, *sentiment analysis* describes a strategy that empowers machines to acquire subjective information, with the greater objective be the digitalisation of human emotion. It falls within the field of research known as natural language processing (NLP)[†], a term first appearing in the 1950's, as computers first began to receive more attention and people wondered whether they could be taught to learn - a classic example of the times being the Turing Test[1] (Recent book covering it's history: [13]). NLP covers a broad spectrum of topics, and can best be summarised as the study of interaction between computers and humans, through the medium of their respective languages. On reporting significant progress in 1954 - successfully translating 60 sentences from Russian language to English language - researchers predicted ([3]) that machine translation would be a solved problem within the decade. Over 60 years later and the problem persists, despite advances in linguistics and computing, and is likely to remain in this state for some time to come (see section 1.3 for more on this).

The field of Sentiment Analysis is constantly evolving, enhancing existing models and creating new ones. This is, firstly, *possible* due in part to new ideas and approaches towards NLP and also because of advancements in the technologies that are the engine behind such the analytical part, i.e. the algorithms. Secondly, the evolution is *necessary* as to keep pace with the target subject. Natural languages are dynamic, changing over time, which makes a *final solution* an unlikely outcome. A model defined today, which can perfectly quantify the sentiment of text (e.g. a tweet), offers little or no guarantee of performing so well next year. New words, expressions and configurations of the two are created on a daily basis, in all natural languages.

[1] Alan Turing introduced this idea in his paper: Computing Machinery and Intelligence[†] [16].

A thorough treatment of the sentiment analysis, its history and development is provided by Liu (2012) [7]. As the inclusion of *sentiment* into financial models is somewhat of a divergence away from pure facts and statistics (at least in its essence), two further related books that explore this realm should be mentioned. The first is *Irrational Exuberance*† [14], by Nobel Prize winning Economist Robert J. Shiller, and the second is *Thinking, Fast and Slow*† by Daniel Kahneman [6], a psychologist who also won the Nobel Prize for economics for his work on Behavioural Finance.

In the framework of behavioural finance, Kahneman introduced a model named *Prospect Theory*, which formed the basis for much research in the intersection between psychology and finance, and a chapter[2] in this book is dedicated to it. As a direct reference to this work, Prospect Theory defines three cognitive features, one of which (*loss aversion*) is directly visible in the sentiment scores and everyday investment decisions. Kahneman succinctly describes this by saying "when directly compared or weighted against each other, losses loom larger than gains". This is evident within the way sentiment scores are computed, where negative words imply larger magnitudes of emotion than their positive counterparts (see footnote Section 1.2 for more detail). In the case of investment decisions things become slightly more complicated, but the idea remains the same; one compares absolute gains and losses with regards to a reference point (e.g. starting capital) to define relative magnitudes. Kahneman mentions that these tendencies occur due to our evolutionary process: "Organisms that treat threats as more urgent than opportunities have a better chance to survive and reproduce".

Shiller's book uses many ideas from Kahneman's works, and addresses more closely the side of finance and how larger market movements may be underpinned by behavioural finance. He gives specific exmaples with stock markets, housing markets and (in the most recent edition of this book, cited above) the bond market.

1.2 Scoring systems

When presented with raw text, the goal of a sentiment analysis algorithm is to assess that text in a way that produces output, which a machine can in turn comprehend, process and use. It does this by using a pre-defined set of rules (a grammar system) and a library of tuples (a dictionary), where each tuple consists of a text string and a corresponding numerical value. Each sentiment analysis model (SAM) has its own grammar system and its own dictionary.

The grammar is what defines the SAM, as it facilitates the interpretation of natural languages (in the case of this study, English) and furthermore acts to translate features of the natural language into a form that machines are able to interpret. This is performed through the quantification of text, assigning scores to words, which is where the dictionary is called upon.

Several dummy examples of word-value tuples are given in Table 1. Eight tuples in total are shown with their output from two different SAMs, who we assume to both produce results on a scale from -5 to $+5$. Intuitively, a negative score implies a negative emotion (or sentiment), while a positive score reflects a positive emotion. Scores close to 0 are to be seen as more neutral. The output contained in the *Score* column highlights similarities and differences between the models as well as several model-specific features.

Looking at the first four scores from **Model 1**, we see some pragmatic ideas have been implemented. `Splendid` definitely has a postive meaning, stronger than simply "good" and so receives a strong positive score. The second term `meadow` on the other hand is difficult to associate with either positive or negative emotions. This is the case for many inanimate objects; it is perhaps impossible to define and assign sentiment scores for words such as "chair", "manufacturing" and "boulder". The word `love` of course returns a positive value, however its common usage (often synonymous with *to like* e.g. "I love pizza") means that the value associated is not as high as one might expect. Furthermore, as discussed in Section 1.1, positive words do no measure as highly on an emotion scale as negative words[3]. Using empirical reasoning like this to adjust scores is very important when analysing social media data as the text is, more often than not, informal. Lastly for

[2] *Thinking, Fast and Slow* ([6] p.278 - 288).

[3] Compare simlar positive and negative words side-by-side to realise this. For example, in the sentence "in my opinion, it really is <*word*>!" replace <*word*> the following word-pairs into sentences and assess your emotional response: [delicious, disgusting], [delightful, sickening] and [beautiful, ugly]. You should notice that the second word in each pair invokes a stronger emotional response than the first.

Model	Word	Score
1	splendid	3
1	meadow	0
1	love	3
1	:)	2
2	love	4
2	:)	0
2	pessimistic	-2
2	extremely	*adaptive*

Table 1: Example of SAM-tuples, where each word is assigned a numerical score.

model 1, the 'happy' *emoji* or *smiley* `:)` has as assigned value that accurately portrays the sentiment. This is again an example of the lexicon in use being modernised to accomdate the evolution of the target content: social media data, including words that were neither present in the early dictionaries created in 1954 nor are in contemporary, conventional dictionaries.

In **Model 2**, the negative word `pessimistic` receives a reasonable score of -2 and compared to **Model 1** has a higher value assigned to `love`. More interestingly, however, it returns a value of zero for the emoji. This shows that this particular SAM does likely not contain tuples within its dictionary to deal with emojis. This point is discussed briefly in Section 1.3. The last word, `extremely`, introduces an interesting case, as its interpretation is bipolar when mapped to emotion. The words up until now were either nouns or adjectives, conveying unambiguous meanings on their own, whereas `extremely` (being and adverb) acts primarily as an *intensifier* of the word that it modifies/describes. For example, compare "extremely satisfied" with "extremely disappointed" - the effect of the adverb increases the magnitude of the emotion, regardless of the nature of that emotion i.e. whether positive or negative. In many of the models, adverbs such as *extremely* are treated as scalars, $\mathbf{s}$, and so, when preceding a word with a sentiment of magnitude $\mathcal{M}$, scale that emotion accordingly: $\mathbf{s} \cdot \mathcal{M}$.

1.3 Difficulties and pitfalls

Sarcasm, irony and many other human emotions are of course extremely difficult to capture without additional information providing the context. This is not a problem specific to machines - humans also often mis-interpret natural languages. For example, if while at the airport I am told that my flight has been cancelled, I may remark "splendid" in a down-beat way. It is clearly a sarcastic remark to the bad news, however the emotion behind the word if given without context is impossible to distinguish, even for a human. Due to such limitations, it must be made clear that the results taken from SAMs cannot be accepted as wholly accurate. The methods involved (described in Section 1.4) are based on good scientific reasoning and research, however also by nature include certain levels heuristics and approximation.

A second limitation (or *feature*, depending on the case at hand) is one already touched upon earlier - that words not included in a dictionary are disregarded. This is indeed the default behaviour of **all** the sentiment models: **when a string is not found within the dictionary, the term is ignored**. This is useful given, the scraped data in this study may contain some words or phrases that are non-sensical. For example, all *hashtags* that remain in the tweets after cleaning are likely non-standard words because they are generally composed of two or more words without spaces. In the example tweet provided in Section `cleaning-tweets`, the hash tag `#trendfollowing` becomes `trendfollowing` after cleaning, which is still not a word that would be found in a dictionary. Unknown words being disregarded is also more favourable than applying a score of zero to them, as that would bias sentiment scores towards zero, with the bias related to the breadth of the dictionary used. The unfortunate aspect of such a model-facet is that, in the particular case of Twitter data, the dictionaries pose a limiting factor. Hashtags, for example, play a large role in the Twitter community,

new ones being created every day which exponential usage. This kind of information could potentially be harnessed to capture short-term trends and information flow, but is alas left untapped with the methods employed in this study.

1.4 Sentiment analysis models

In this section we outline the five models that were used to score each individual tweet that was obtained via the Twitter mining process described in Section `iterative-scraping`. Each of the models approaches sentiment analysis from a slightly different angle. However, as this study is primarily focused on the implementation of sentiment analysis within a financial context, detailed descriptions of each of the models and their corresponding algorithms are not provided, rather links to the literature.

Many of the models were written to return an integer value, however the underlying code[4] of the SAMs was altered at their final step to simply return the decimal value, if possible. While integer values may be more easily interpreted when making comparisons between invidual tweets, this was not the use-case for this study. As the sentiment data from each individual tweet was eventually aggregated with others from the same day, it made sense that each tweet held its value in its most precise form, i.e. the raw decimal value. Using the decimal values provides a final average for any specific day that does not compound any rounding errors. Furthermore, the statistical methods used (as described in Section `comp-boosting`) are not confined to using integers integers.

1.4.1 EmoLex

This model, formally named the NRC Word-Emotion Association Lexicon† is an open source project that supports 20 natural languages (at the time of writing). In their related publications [10], the authors Mohammad and Turney develop a robust framework to produce accurate assignment of sentiment scores to words and terms. This first uses a level of abstraction when assigning scores to words, where the *sense* of the word is decomposed into multiple classes. EmoLex asks the contributors, who **manually** annotate the words, to assess each individual word or term according to *eight* different axes of emotion; these which are displayed in the outermost ring in Figure 1[5]. Derived by psychologist Plutchik, [12], this *wheel of emotions* defines a scale to reflect the magnitude of each of the eight compound emotions. The innermost ring at the centre of the wheel being the strongest level of a given emotion, the outermost ring the weakest. The eight scores assigned to each word or term are then aggregated by the sentiment analysis model used, meaning that only one output value is given for each word. the final results is dependent on whether the total positive emotion is greater than the negative emotion in each case, and vice versa. For more information, please refer to the above referenced literature.

The second of the two fundamental ideas behind this model is that the dictionary used should not be a limiting factor for the analysis of sentiment within text. What this translates into, practically speaking, is an extremely large dictionary of emotions being desired. This was achieved through applying a separate methodology, namely *crowd-sourcing*†, (comparable to *crowd-funding*) a term springing mainly from online collaborations, whereby many people each make a small contribution to a large project. In this case, people were paid a small amount for each pre-defined batch of words that they manually assigned sentiment scores to.

1.4.2 Sentiment140

This SAM is based on a lexicon that originates from the creators of EmoLex[6], but differs to it by the method of its creation, in that an **automated** process was used (detailed below). Such a process has the benefit of being able to assess and learn from many more words than were used for the manually assigned scores of EmoLex. The algorithm is explained in detail within the literature [9]; here we provide an explanation only of the essential points.

From a corpus of tweets that each contained an emoji, all possible unigram (one-word) and bigram (two-word) combinations were made from the words contained in that tweet, any non-intelligible cases being

[4]Provided by Matheus Araújo† from iFeel - the online sentiment analysis application†

[5]The image file is taken from the Wikemedia Commons (last visited: 15th March 2016)†.

[6]There are numerous additional dictionaries defined by EmoLex creator Saif Mohammed†.

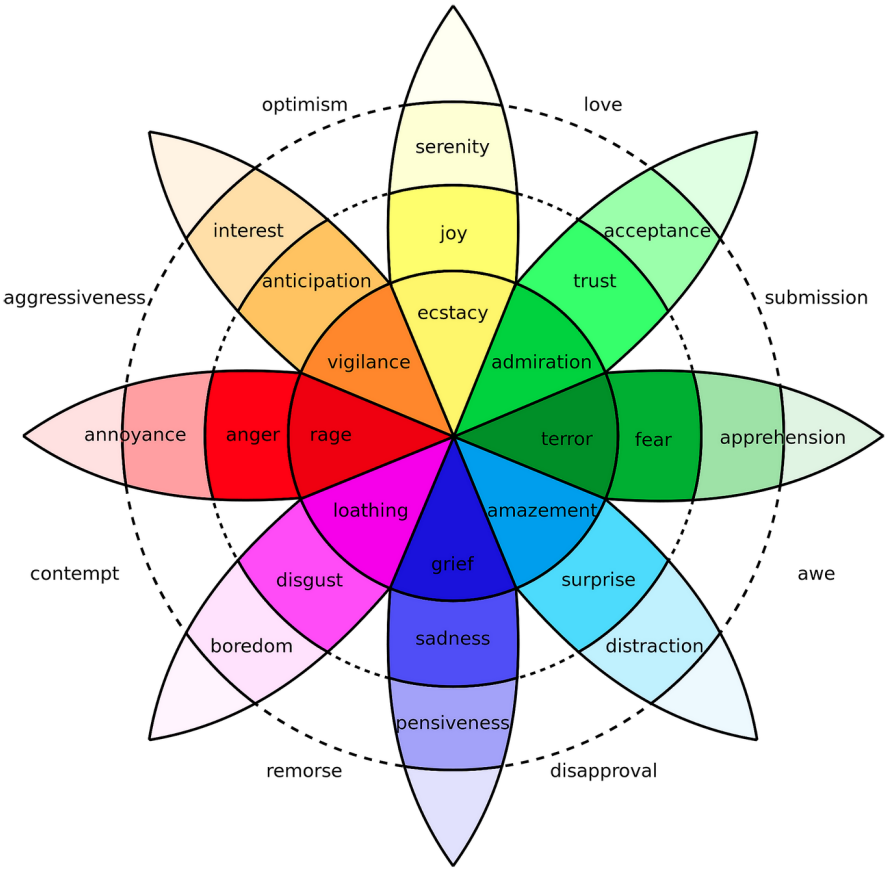

Figure 1: Plutchik's *Wheel of Emotion* - eight axes define eight basic emotions, neighbouring pairs of which are used to derive eight *compound emotions*, given between each basic pair on the outskirts of the wheel. The colour-intensity of each axis signifies the intensity of that basic emotion. Darker colours, moving towards the centre of the wheel, represent larger magnitudes of emotion.

subsequently removed via a customised filter. Each unigram and bigram was then assigned a value of -5, 0 or $+5$ depending on the related emoji's predefined score. In total 1.6 million data points were created, and assigned value using this pre-definined dictionary of emojis. The scores assigned to the uni- and bigrams were then used to construct a new dictionary, which can itself be used to evaluate new tweets. The scores of all contained uni- and bigrams within a single tweet are summed to provide the final result.

That the output from the Sentiment140 model can clearly be in the form of large numbers (given each dictionary entry corresponds to either -5, 0 or $+5$) is irrelevant for this study, as all data was later normalised to a smaller range - see Section `rescaling-sent` for more details.

1.4.3 SentiStrength

This SAM was selected in order to add further diversity to the methodologies used within the five models. The SentiStrength method [15], available in fifteen languages[†], presents a novel method of analysing sentiment stemming from concepts within the field of psychology - with the notion that humans experience more than one emotion at the same time, with the larger emotions overwhelming the smaller emotions. Expressed differently, each time a person experiences a feeling towards a given event or object, it can be measured on more than one axis, in fact the model outlines three distinct axes. Given an event, defined as anything that may evoke an emotion, three levels of emotion are mapped to these three independent axes, where:

Axis 1: describes positivity
Axis 2: describes negativity
Axis 3: describes neutrality

This approach decomposes human emotion, using parameterization to describe something we often experience, and would indeed describe, as one single emotion or feeling. For the purposes of this study, a binary response was chosen, leaving out neutrality. This is because a sense of neutrality can be inferred from the magnitude of the first two axes. If low scores are obtained from both positive and negative feelings regarding an event, then the neutrality score must be relatively high. The results from the binary positive-negative scale were obtained and later combined in the modelling preparation steps outlined in Section `rescaling-sent`.

1.4.4 VADER and VADER AFinn

Valence **A**ware **D**ictionary for s**E**ntiment **R**easoning (VADER) - is a parsimonious rule-based model for sentiment analysis of social media [4]. The creators constructed a dictionary targeted towards micro-blogging content e.g. Twitter, via a combination of both quantitative and qualitative methods. At the heart of the algorithm are five simple rules that describe both grammatical and syntactical conventions that are honed to detect markers of sentiment (or *Valence* in VADER terminology) with text. The authors claim that their model not only improves many well-established benchmarks that use more complicated methods, but also outperforms individual human raters. A useful facet of being derived solely using parsimonious techniques is that the model can be generalised quite well into other domains beyond Twitter.

VADER not only includes emojis, but also incorporates common slang terms (e.g. "nah" and "meh"), but also acronyms and initialisms, such as the widely used terms "LOL" and "WTF". Like many other creators of sentiment analysis dictionaries, the authors made use of Amazon Mechanical Turk[†7], meaning each of their words was **manually** assigned a value by a human. The authors also created some extra steps to this process (detailed in the above referenced literature) to ensure that the evaluation of all tweets was performed to the highest standard, to produce what they term a *gold-standard* lexicon. This is a crucial point, considering the model itself uses only a small number of rules to classify tweets as positive or negative.

VADER AFinn, built upon the VADER lexicon (the AFinn suffix derived from the author's name) defines yet another lexicon, which includes higher amounts of slang and even *vulgar slang*[†] - words that not many lexicons include, but are commonly seen on micro-blogging websites. In the related literature, [11], the author shows several examples where the inclusion of such words does indeed more accurately reflect the true sentiment of the text. Syntactical analysis is also performed just as in VADER, where e.g. the words "but" is used in the quantification step as a contrastive conjunction. This means that "but" makes a contrast where the text that follows it reflects a stronger sentiment than the text preceding it. The tweet text before is reduced in intensity and the text following is increased in intensity to take this into account.

1.4.5 Example analyses

All tweets taken and cleaned from Twitter Mining saved in text files with one tweet per line read through each sentiment model, outputs returned in one table. Table 2 contains scores provided by the different SAMs for several sample tweets. Without describing in great deal how the results differ, it is clear that the described features of each model do have an impact when analysing informal text, as found within the Twitter data. Worth noting are (1) the binary output from the SentiStrength model, the first results representing how *negative* the tweet is and the second how *positive*, and (2) the zero value for VADER in Example 3. The two values of both halves of the sentence in `Example 3` balance out to create an overall zero outcome, which is not observed in the other models. The results for the same tweet, being relatively negative for all other models, reflects the property in the model described in Section 1.4.4, using conjuntions such as "but" to accordingly weight the main and subordinate clauses being connected.

[7] A system whereby *employers* create tasks to be completed, (usually repetitive and not requiring any introduction) that *employees* can slowly work through at any given time. It means humans perform the work and is an extension of crowd sourcing, mentioned in section 1.4.1.

ID	Tweet text	EmoLex	Sentiment140	SentiStrength	VADER	V. AFinn
1	I love you :-) LOL	+ 0.7	+ 0.6	(− 1, + 4)	+ 0.8	+ 1.0
2	I hate you :-(	− 0.9	− 0.9	(− 5, + 1)	− 0.8	− 1.0
3	I like cats, but hate dogs	− 0.3	− 0.1	(− 4, + 3)	0	− 0.5

Table 2: Three example tweets with sentiment scores from each model.

1.5 Applications in finance

Here we give three examples of related works, which display how sentiment analysis can be applied within a financial setting, and how it indeed was at different points over the past decade.

1.5.1 In 2006

Baker and Wurgler, [1], created and implemented a sentiment driven model to investigate cross-sections of returns. Up until that point in time, classical financial theory was used to explain how the diversification methods among rational investors leads to an equilibrium in the market, which precisely portrays all rationally discounted cash flows. This general statement is supported by the Efficient Market Hypothesis (EMH), defined by Fama [8] as follows: "*prices fully and instantaneously reflect all publicly available information*"[8]. The work carried out by Baker and Wurgler involved defining three *proxies* to investor sentiment (there was no social media data readily available at the time), which were: the book-to-market ratio, external financing and sales growth. The authors recognised that these were not direct indicators of sentiment, hence naming them proxies, and so took the first principal component of the data set as the final variable. They were able to make two statements from their results regarding the reurns of certain categories of stocks[9]:

1. when beginning-of-period proxies for sentiment are low, subsequent returns are relatively high, but
2. when sentiment is high, relatively low subsequent returns may be expected.

1.5.2 In 2010

Bollen *et al.* [2] show how Twitter sentiment can be used to predict stock markets movements using sentiment results from two different SAMs (OpinionFinder[†] and Google-Profile of Mood States (GPOMS)[10]). The GPOMS model is similar to the EmoLex model described in Section 1.4.1, defining various scales of emotion. Using one of these elements along with historical market data within a self-organising fuzzy neural network model, the authors were able to make predictions regarding market movement with accuracies greater than 80 %. The work was influenced by ideas stemming from behavioural finance, the authors being able to gain Kahneman's input (see Section 1.1) regarding their model and utilising notions from his Prospect Theory.

1.5.3 In 2012 and beyond

The last example is that of major ongoing project MarketPsych[†], an index of sentiment compiled by Thomson Reuters over a vast number of industries. Many sources of data are used, but are able to be placed into three categories: high-frequency data in the form of social media messages from 1998 to present, medium frequency data obtained through the trawling (web-scraping) of countless internet news sites and live, low frequency data obtained directly from Retures itself. The algorithms that follow are very similar to those employed in this study, with all the information boiling down to numerical indicators. These are what form the Thomson Reuters MarketPsych Indices (TRMI), which had modest beginnings, only covering several

[8] Alternatively formulated from the perspective of arbitrage by Jensen [5]: "A market is efficient with respect to an information set, if it is impossible to make economic profits by trading on the basis of this information".

[9] Small stocks, young stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme growth stocks, and distressed stocks.

[10] This is a modified version of a well-known psycological test[†], which was adapted by the authors for use with Google data.

general asset classes (such as agriculture and energy), but now includes many thousands of indicators going as far as focusing on specific companies. One great area of success shown by TRMI has been the recognition and prediction of market bubbles[†]. Arguing that bubbles are no longer only sparse events, they instead - with the rapidly growing international connectivity and corresponding currents of money within financial markets - additionally describe short-lived *specultive mania*, which the real-time analysis of sentiment data allows investors to track.

References

[1] MALCOLM BAKER and JEFFREY WURGLER. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.

[2] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[3] Leon E Dostert. The georgetown-ibm experiment. *1955). Machine translation of languages. John Wiley & Sons, New York*, pages 124–135, 1955.

[4] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[5] Michael C Jensen. Some anomalous evidence regarding market efficiency. *Journal of financial economics*, 6(2/3):95–101, 1978.

[6] Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

[7] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[8] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

[9] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[10] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.

[11] Finn Arup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May 2011.

[12] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.

[13] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. *The Turing Test: The Elusive Standard of Artificial Intelligence*, chapter Turing Test: 50 Years Later, pages 23–78. Springer Netherlands, Dordrecht, 2003.

[14] Robert J Shiller. *Irrational Exuberance*. Princeton university press, 2015.

[15] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010.

[16] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.