Ein Konzept für die Klassifizierung subjektiver Sicherheit in Tweets

Kristian Rother, Inga Karl, Simon Nestler

Forschungsgruppe Mensch-Computer-Interaktion, Hochschule Hamm-Lippstadt

Zusammenfassung

In diesem Beitrag wird die erste Iteration eines Prozesses zur Konzeption einer Annotationsrichtlinie zur Klassifizierung von Tweets hinsichtlich des in ihnen ausgedrückten subjektiven Sicherheitsgefühls dargelegt. Basierend auf einer initialen, rudimentären Annotationsrichtlinie wurden Tweets von vier Annotatoren klassifiziert. Diese Klassifizierung wurde zu einem späteren Zeitpunkt wiederholt und die Annotationen wurden hinsichtlich der Interrater- und Intrarater-Reliabilität untersucht. Anhand qualitativer Interviews wurden Handlungsempfehlungen für die Überarbeitung der Annotationsrichtlinie abgeleitet.

1 Einleitung

Eine Möglichkeit zur Erfassung des subjektiven Sicherheitsgefühls der Bevölkerung im Kontext der Mensch-Computer-Interaktion ist die Analyse von verbreiteten Inhalten in Sozialen Medien. Im Rahmen des Forschungsprojektes INTERKOM[1] in der zivilen Sicherheitsforschung wird ein System entwickelt, mit dem die Inhalte aus Sozialen Medien analysiert und hinsichtlich der subjektiven Sicherheitswahrnehmung bewertet werden sollen. Eine der notwendigen Teilaufgaben für ein solches System ist die automatisierte Klassifizierung von verbreiteten Diskussionsinhalten in Sozialen Medien am Beispiel von Twitter. Die Klassifizierung erfolgt hinsichtlich des subjektiven Sicherheitsgefühls, welches in den Kurznachrichten (Tweets) ausgedrückt wird.

Um Tweets computerbasiert mithilfe von überwachtem Lernen klassifizieren zu können, wird ein Goldstandard benötigt (Raykar et al., 2009). Bisher gibt es keinen Goldstandard für die Erhebung des subjektiven Sicherheitsgefühls, welches in deutschsprachigen Tweets ausgedrückt wird. Um einen solchen Goldstandard zu schaffen, ist eine Richtlinie zur Klassifizierung der Tweets notwendig. In diesem Beitrag wird die erste Iteration eines Prozesses zur Konzeption dieser Annotationsrichtlinie vorgestellt. Entsprechend lautet die Forschungsfrage: Wie lässt sich eine wissenschaftlich fundierte Annotationsrichtlinie für das in deutschsprachigen Tweets ausgedrückte subjektive Sicherheitsgefühl entwickeln?

[1] BMBF-Förderkennzeichen 13N1005, 01/2014 – 12/2016

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2016 in
B. Weyers, A. Dittmar (Hrsg.):
Mensch und Computer 2016 – Workshopbeiträge, 4. - 7. September 2016, Aachen.

http://dx.doi.org/10.18420/muc2016-ws01-0007

2 Verwandte Arbeiten

Bisherige Ansätze zur Analyse von Krisensituationen anhand von Twitter, beispielsweise im Kontext des Tsunamis in Japan (Acar & Muraki, 2011), der Flut in Queensland (Bruns et al., 2012) und des Erdbebens in Chile (Mendoza et al., 2010) beschränken sich auf den objektiven Krisenkontext und untersuchen die Art der Tweets, die in Krisensituationen gesendet werden, oder wie sich diese verbreiten. Naaman, Boase und Lai (2010) haben inhaltsbasiert unterschiedliche Kategorien von Tweets abgeleitet. Die Kategorie „Suche nach Informationen, Hilfe oder Antworten“ wird im Kontext der Black Saturday Feuer in Australien als besonders krisenrelevant bezeichnet (Sinnappan et al., 2010).

Im Gegensatz dazu liegt der Fokus bei der Sentiment Analyse auf in Texten zum Ausdruck gebrachten Meinungen, Gefühlen und Subjektivität (Pang & Lee, 2008). Die Verwendung von Twitter als Korpus für die Sentiment Analyse zur Klassifizierung in „positiv“, „negativ“ und „neutral“ mittels überwachtem Lernen wurde von Pak und Paroubek (2010) sowie Go und Kollegen (2009) dargelegt. Saif und Kollegen (2013) erweitern die Sentiment Klassifizierung von Tweets als Ganzes um Annotationen für in den Tweets enthaltene Entitäten. Eine einfache Klassifizierung in Gefühle basierend auf Schlüsselwörtern findet sich im Kontext des Terrorismus bei Cheong und Lee (2011).

Wiebe und Kollegen (1999) stellen einen Prozess vor, um anhand einer ersten Annotationsrichtlinie und statistischen Tests eine verbesserte Annotationsrichtlinie abzuleiten. Ein weiterer Prozess zur iterativen Verbesserung von Annotationsrichtlinien findet sich in den ersten beiden Phasen (Modellierung und Annotation) des MATTER-Zyklus, welche wiederum den MAMA-Zyklus beinhalten (Pustejovsky & Stubbs, 2012, S. 23–29). In Anlehnung an den MAMA-Zyklus sowie an Wiebe und Kollegen (1999) soll die Entwicklung der Annotationsrichtlinie daher in einem iterativen Prozess erfolgen, wobei diese in jeder Iteration getestet und angepasst wird, bis eine standardisierte Annotationsrichtlinie vorliegt.

3 Methode

Der Gesamtprozess zur Entwicklung des Goldstandards ist in Abbildung 1 zusammengefasst. Dieser Beitrag beschreibt einen Durchlauf des iterativen Workshop-Annotation-Evaluation Zyklus. Dieser Zyklus wird solange durchlaufen, bis im Rahmen der Evaluation gezeigt werden kann, dass die Annotationsrichtlinie den Anforderungen entspricht. Im ersten Schritt des Zyklus wurde zunächst ein Kickoff-Workshop der Forschungsgruppe Mensch-Computer-Interaktion der Hochschule Hamm-Lippstadt dazu genutzt, die erste Annotationsrichtlinie zu entwickeln.

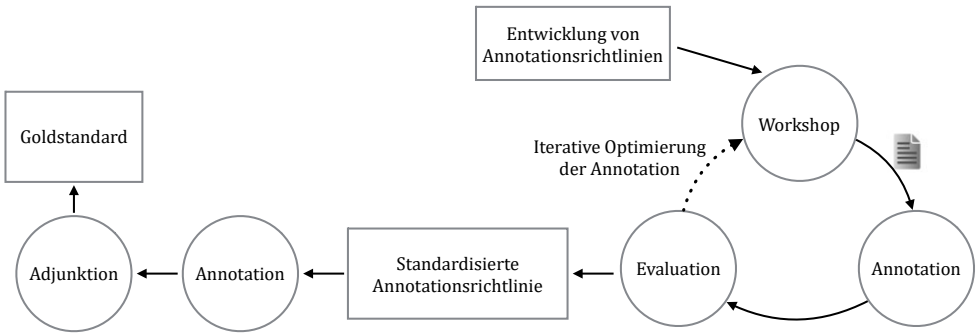

Abbildung 1: Prozess für die Entwicklung eines Goldstandards in Anlehnung an Wiebe et al. (1999) und Pustejovsky und Stubbs (2012, S. 23–29)

Die Forscher einigten sich darauf, die initiale Annotationsrichtlinie möglichst einfach und rudimentär zu halten, sodass die ersten Ergebnisse als Bestandsaufnahme dienen und so das relevante Verbesserungspotential bestimmt werden kann. Die entwickelte Annotationsrichtlinie sah wie folgt aus:

Ziel: Das Ziel der Annotation ist es, Tweets hinsichtlich der gefühlten Sicherheit des Verfassers der Nachricht zu klassifizieren.

Klassenbezeichnungen: Es stehen drei Klassenbezeichnungen zur Auswahl. „Gefühlte Sicherheit" bedeutet, der Tweet drückt ein Gefühl von Sicherheit aus. „Gefühlte Unsicherheit" bedeutet, der Tweet drückt ein Gefühl von Unsicherheit aus und „Neutral" bedeutet der Tweet drückt weder ein Gefühl von Sicherheit noch ein Gefühl von Unsicherheit aus.

Durchführung: Die Klassifizierung bezieht sich immer auf den gesamten Tweet. Betrachten Sie immer nur einen Tweet. Beginnen Sie beim ersten (obersten) Tweet und wählen Sie die Ihrer Meinung nach passende Klassenbezeichnung im Dropdownmenü aus. Fahren Sie mit dem nächsten Tweet fort, bis Sie am Ende des Dokumentes angekommen sind.

Nachdem die initiale Annotationsrichtlinie vorliegt, entspricht der Zyklus einem fortlaufenden sequentiellen *Mixed Methods Design* (Ivankova et al., 2006), d.h. zeitlich geht die quantitative Datenerhebung (Abb. 1: Annotation und Evaluation) der qualitativen Untersuchung (Abb. 1: Evaluation und Workshop) voraus. Die quantitativen Daten wurden gewonnen, indem mehrere Annotatoren auf Basis der ersten Annotationsrichtlinie Tweets hinsichtlich der gefühlten Sicherheit klassifiziert haben. Die so gewonnenen Daten wurden nach auffälligen Mustern untersucht. Basierend auf den Ergebnissen dieser Untersuchung wurden qualitative Interviews mit den Annotatoren geführt, um ihren Klassifizierungsprozess besser zu verstehen und zu erklären.

3.1 Materialien

Für die Analysen wurden Tweets aus einer Datenbank entnommen, die im Rahmen des Forschungsprojektes angelegt wurde und seit Okt. 2015 Tweets mithilfe der von Twitter bereit-

gestellten öffentlichen Streaming-API[2] sammelt. Im Kontext des Projektes werden die Tweets mithilfe von 88 domänenspezifischen Schlüsselwörtern (u.a. Anschlag, Polizei, ansteckend usw.) gesammelt, um zu erreichen, dass der Datensatz möglichst kontextrelevante Inhalte aufweist (d.h. deutsche Tweets mit Krisenkontext). Diese Schlüsselwörter wurden durch Experten auf Basis eines in Entwicklung befindlichen standardisierten Fragebogens[3] zum subjektiven Sicherheitsgefühl der Bevölkerung abgeleitet. Für die vorgestellte Studie wurden aus der Datenbank insgesamt 250 Tweets als Trainingsdaten ausgelesen. Aus den Tweets wurde mit einem Python-Skript eine LibreOffice Calc Tabelle für jeden Annotator generiert. Die Inhalte der einzelnen Tweets wurden in der ersten Spalte unter der Überschrift „Tweet“ ohne Autorenangaben gespeichert. Die zweite Spalte, mit der Überschrift „Klassifizierung“ enthielt ein Auswahlmenü mit den drei Klassenbezeichnungen. Aufgrund der Zielsetzung, die Tweets hinsichtlich der gefühlten Sicherheit des Verfassers der Nachricht zu klassifizieren und weil letztendlich das subjektive Sicherheitsgefühl der Bürger untersucht werden soll, wurden Tweets, die direkt von einer Behörde stammen, in diesem Datensatz nicht verwendet. Retweets, d.h. das (meist unkommentierte) Weiterverbreiten einer Behördennachricht durch den Bürger, wurden hingegen als relevant betrachtet und aufgenommen. Ob es sich um einen Retweet handelt ließ sich den Metadaten der Tweets entnehmen.

3.2 Durchführung

Die Klassifizierung der gesammelten Tweets wurde von insgesamt vier Personen (A-D) durchgeführt. Die Annotatoren A, B und C waren grob mit der Domäne Sicherheitsforschung vertraut und Annotator D hatte ein tieferes Domänenwissen. Alle vier Annotatoren waren männlich und zwischen 21 und 33 Jahre alt. Es wurden insgesamt zwei Erhebungen durchgeführt, wobei die Annotatoren nicht wussten, dass ein weiterer Durchlauf geplant war. Beim ersten Durchlauf im November 2015 wurden die Annotatoren in die Verwendung des LibreOffice Calc Dokuments eingeführt. Dabei wurden ihnen die Annotationsrichtlinie sowie die Auswahl der gewünschten Klassenbezeichnungen für einen Tweet erklärt. Dann konnten die Annotatoren unbeobachtet die Tabelle mit den Tweets bearbeiten. Der zweite Durchlauf im Januar 2016 verlief analog zum ersten, allerdings wurden die Annotatoren anschließend gebeten sich zu notieren wie sie bei der Klassifizierung vorgegangen sind und was ihnen leicht oder schwer gefallen ist.

4 Ergebnisse

Die Ergebnisse beziehen sich auf die Annotation der Tweets durch die vier Teilnehmer. Die Auswertung der Klassifizierung erfolgte mit dem R-Paket irr (Gamer et al., 2012). Um zu

[2] Informationen zur API finden sich unter https://dev.twitter.com/streaming/reference/post/statuses/filter

[3] Die verwendete vorab Version des Fragebogen stammt aus privater Korrespondenz mit dem Universitätsklinikum Ulm.

untersuchen wie unterschiedlich die vier Annotatoren jeweils einen Tweet bewertet haben, wurde die Interrater-Reliabilität (Fleiss' Kappa) mit der Funktion *kappam.fleiss* ermittelt. Um zu untersuchen wie unterschiedlich einzelne Tweets von den jeweiligen Annotatoren an den beiden Erhebungszeitpunkten eingestuft wurden, wurde die Intrarater-Reliabilität (Cohen's Kappa) mit der Funktion *kappa2* ermittelt. Zur Beurteilung der Ergebnisse werden die sechs Bewertungen der κ-Werte von Landis und Koch (1977) verwendet: *schlechte Übereinstimmung* (<0), *etwas Übereinstimmung* (0 bis 0.20), *ausreichende Übereinstimmung* (0.21 bis 0.40), *mittelmäßige Übereinstimmung* (0.41 bis 0.60), *beachtliche Übereinstimmung* (0.61 bis 0.80) und *(fast) vollkommene Übereinstimmung* (0.81 bis 1).

4.1 Klassifizierung der Tweets

Tabelle 1 zeigt die Klassifizierung der Tweets an den beiden Erhebungszeiträumen. Für die Interrater-Reliabilität der vier Annotatoren ergab sich beim ersten Durchgang ein *Fleiss' Kappa* von $\kappa = 0.48$ ($z = 21.60$; $p < 0.01$) und beim zweiten Durchgang ein *Fleiss' Kappa* von $\kappa = 0.53$ ($z = 23.70$; $p < 0.01$).

Annotator	**Gefühlte Sicherheit**		**Gefühlte Unsicherheit**		**Neutral**	
	Nov 2015	Jan 2016	Nov 2015	Jan 2016	Nov 2015	Jan 2016
A	18	15	60	60	172	175
B	8	8	64	68	178	174
C	18	3	74	32	158	215
D	4	14	29	53	217	183

Tabelle 1: Klassifizierung der 250 Tweets an den beiden Erhebungszeiträumen

Die Intrarater-Reliabilität für die jeweiligen Annotatoren über zwei Zeitperioden hinweg ist in Tabelle 2 dargestellt.

	Ungewichtet			Gewichtet mit g = <0,1,2>		
Annotator	**Cohen's Kappa**	**z-Wert**	**p-Wert**	**Cohen's Kappa**	**z-Wert**	**p-Wert**
A	0.68	12.90	<0,01	0.69	13.10	<0.01
B	0.82	14.30	<0,01	0.78	13.90	<0.01
C	0.25	4.23	<0,01	0.22	3.88	<0.01
D	0.61	11.80	<0,01	0.58	11.10	<0.01

Tabelle 2: Intrarater-Reliabilität über zwei Zeitperioden

4.2 Konfliktäre Annotationen

Im ersten Durchgang kam es insgesamt bei sechs der 250 Tweets (2.4 %) zu konfliktären Annotation d.h. mindestens ein Annotator hat den Tweet mit GS klassifiziert und ein anderer Annotator hat denselben Tweet mit GU klassifiziert. Im zweiten Durchgang war dies bei sieben einzigartigen Tweets von 250 (2.8 %) der Fall. Insgesamt gab es drei Tweets (1.2 %), die in beiden Durchgängen konfliktär annotiert wurden. Diese sind im Folgenden aufgeführt.

- RT @BILD_Muenchen: Mit #Fahndungsfoto: #Polizei sucht brutalen #Schläger https://t.co/VWzE3vBRNU https://t.co/DzqJIj316v
- Öffentlichkeitsfahndung: Polizei München sucht drei mutmaßliche Schläger https://t.co/NkzOm9QNz0
- Mannheim-Rheinau – Polizei warnt vor unlauteren Schmuckankäufen – Zeugen und Geschädigte gesucht! #Ludwigshafen https://t.co/RRgO1sRtxY

5 Diskussion

Die Ergebnisse zeigen eine *mittelmäßige Übereinstimmung* der Interrater-Reliabilität in beiden Durchgängen. Es wird daher vermutet, dass es grundsätzlich möglich ist das Sicherheitsgefühl von deutschsprachigen Tweets auf der Dokumentenebene zu klassifizieren. Da die Annotatoren nur mit einer rudimentären Annotationsrichtlinie gearbeitet haben, ist eine Verbesserung der Reliabilität für weitere Vorgänge zu erwarten. Die gewichtete Intrarater-Reliabilität zeigt unterschiedliche Übereinstimmungen der Bewertungen der einzelnen Annotatoren. Die Annotatoren A und B liegen dabei innerhalb der Richtwerte für *beachtliche Übereinstimmung* und für Annotator D innerhalb der Richtwerte für *mittelmäßige Übereinstimmung.* Die Intrarater-Reliabilität von Annotator C liegt am unteren Ende der Richtwerte für *ausreichende Übereinstimmung.* Die qualitativen Interviews haben ergeben, dass Annotator C im Vergleich zu den anderen nach der ersten Erhebung bereits über den Annotationsprozess nachgedacht und „quasi den eigenen Prozess angepasst“ hat. Das Infragestellen der Klassifizierung verdeutlicht, dass die Richtlinie so angepasst werden muss, dass die Annotatoren weniger Interpretationen der Richtlinie vornehmen können. Somit ist anzunehmen, dass eine bessere Annotationsrichtlinie dazu führen wird, dass sich die Intrarater-Reliabilität für alle Annotatoren in weiteren Iterationen verbessern wird.

Im Bezug auf die konfliktären Tweets kann vermutet werden, dass die persönliche Wahrnehmung der Inhalte durch die Annotatoren eine starke Rolle bei der Klassifizierung gespielt haben. In diesem Zusammenhang lassen sich aus den Ergebnissen der qualitativen Befragung erste Handlungsempfehlungen für Erweiterungen der Annotationsrichtlinie ableiten, um konfliktäre Einstufungen in weiteren Iterationen zu vermeiden und weniger Spielraum für eigene subjektive Meinungen der Annotatoren zu geben:

- Beispiele für Tweets zu jeder Klassenbezeichnung („Beispiele wären gut gewesen.“)
- Erläuterung des Umgangs mit Retweets („Retweet also mit RT. Das muss ja besonders wichtig sein.“), unvollständigen Tweets („Unvollständige Tweets bewerte ich eher mit

neutral.“) und Tweets, die Links enthalten („Nachrichten mit Links sind schwierig einzuordnen.“) oder Teil eines Dialogs sind („Manchmal sieht man halt nicht, wies weitergeht.“)

- Genauere Erklärungen zu @ und # und wie diese zu Beurteile sind („Ich schaue ob es ein @ oder # gibt was über den Kontext Ausschluss geben kann. Also wenn da jetzt z.B. was über Schüsse steht aber #callofduty dann ist das neutral.“)
- Eine Identifizierung von offiziellen Accounts, insbesondere BOS und Nachrichtenagenturen („Wenn sowas wie @polizei ich glaube Baden-Württemberg dann sieht man ja, dass es von einer Behörde kommt [...] das ist dann offiziell.“)
- Fahndungsmitteilungen erfordern besondere Erklärungen, da sie oft konfliktär eingestuft wurden („Wenn nach Tätern gesucht wird ist das immer so ein gemischtes Ding [...]“. „Fahndungserfolge bedeuten Sicherheit.“)
- Der Bezug zu Deutschland muss genauer erläutert werden („Wäre ein Flugzeugabsturz in Australien jetzt Unsicherheit? Eigentlich ja nicht weil kein Bezug zu Deutschland […] also weiß ich jetzt nicht.“)

6 Ausblick

Ziel dieser Arbeit war es einen Prozess zur Konzeption einer Annotationsrichtlinie zur Klassifizierung von Tweets hinsichtlich des in ihnen ausgedrückten subjektiven Sicherheitsgefühls darzustellen. Es wurde gezeigt, dass bereits mit einer sehr einfachen Annotationsrichtlinie eine *mittelmäßige Übereinstimmung* bei der Klassifizierung erzielt werden kann. Dieser Wert liefert einen ersten Richtwert für zukünftige Verbesserungen und einen Anhaltspunkt, ob die getroffenen Maßnahmen die Reliabilität im Vergleich zur rudimentären Annotationsrichtlinie verbessern. Für die nächste Iteration wird eine Steigerung der Interrater-Reliabilität und der Intrarater-Reliabilität auf mindestens *beträchtliche Übereinstimmung* angestrebt. Als Ergebnis der hier vorgestellten ersten Iteration konnten erste Erweiterungen der Annotationsrichtlinie abgeleitet werden.

In Anlehnung an die Identifikation von Sarkasmus in Tweets nach González-Ibáñez et al. (2011) wäre es denkbar, dass es auch so etwas wie #unsicher gibt. Als Alternative zur Annotation können die Autoren der Tweets direkt befragt werden, ob sie sich zum Zeitpunkt der Erstellung der Nachricht unsicher gefühlt haben (Acar & Muraki, 2011). Sobald eine *beträchtliche Übereinstimmung* erzielt wird, kann durch Adjunktion ein Goldstandard (vgl. Abb. 1) geschaffen werden. Nachdem der Goldstandard geschaffen wurde, können unterschiedliche Algorithmen des überwachten Lernens systematisch verglichen werden.

Danksagungen

Dieser Beitrag wurde durch das Bundesministerium für Bildung und Forschung (BMBF) als Teil des Projektes INTERKOM (FK: 13N1005, 01/2014 – 12/2016) gefördert. Wir bedanken uns bei Marius Gördes, Kai Sablowski und Dennis Ziebart für ihre Unterstützung.

Literaturverzeichnis

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, *7*(3), 392–402.

Bruns, A., Burgess, J. E., Crawford, K., & Shaw, F. (2012). # qldfloods and@ QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods.

Cheong, M., & Lee, V. C. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, *13*(1), 45–59.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. Retrieved from http://CRAN.R-project.org/package=irr

Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. *Entropy*, *17*.

Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, *18*(1), 3–20.

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

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? In *Proceedings of the first workshop on social media analytics* (pp. 71–79). ACM.

Naaman, M., Boase, J., & Lai, C.-H. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 189–192). ACM.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, pp. 1320–1326).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning*. O'Reilly Media, Inc.

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., … Moy, L. (2009). Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning* (pp. 889–896). ACM.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Sinnappan, S., Farrell, C., & Stewart, E. (2010). Priceless tweets! A study on Twitter messages posted during crisis: Black Saturday. *ACIS 2010 Proceedings*, *39*

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and Use of a Gold-standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246–253). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1034678.1034721